Chapter 1

Finite Difference Solution of Linear Second Order Elliptic Partial Differential Equations

1.1 Derivation of the equation

Elliptic partial differential equations can be generally obtained from time dependent problems considering the so-called *stationary case*. In the stationary case the solution varies only with the spatial coordinates and not with time. If the solution is steady in time, then the time derivative term is equal to zero.

1.1.1. Example. Let us consider the one-dimensional heat conduction problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0,1);$$

$$u(t,0) = 0,\ u(t,1) = 1,\ t \geq 0;$$

$$u(0,x) = u_0(x), \quad x \in [0,1].$$

If the initial temperature is described by the linear function $u_0(x) = x$, $x \in [0,1]$, then the solution of the problem is $u(t,x) = x$, that is the temperature is independent of time. This solution is called the stationary solution of the problem. The stationary solution can be obtained by setting $\partial u/\partial t = 0$ and solving the ordinary differential equation

$$u'' = 0, \quad x \in (0,1);$$

$$u(0) = 0,\ u(1) = 1.$$

For other initial functions, say $u_0(x) = x^2$, the temperature will "tend" to the stationary solution. □

The general form of linear second order elliptic partial differential equations is

$$\mathrm{div}(D\ \mathrm{grad} u) + f = 0 \ \text{ in } \Omega, \tag{1.1.1}$$

where functions $D = D(\mathbf{x})$ (coefficient of heat conduction) and $f = f(t, \mathbf{x})$ (the heat source term) are known, and we have to determine the function $u = u(t, \mathbf{x})$. (Function u is supposed to be continuous on $\bar{\Omega}$ and differentiable in Ω.) In order to obtain a well-posed problem, we generally prescribe so-called *boundary conditions* on the boundary (denoted by $\partial\Omega$) of the domain Ω. There are three different types of boundary conditions.

- Dirichlet boundary condition. The function u is prescribed on the boundary, that is $u(t, \mathbf{x}) = g_1(\mathbf{x})$, $\mathbf{x} \in \partial\Omega$ (g_1 is a given function).
- Neumann boundary condition. The (normal) derivative of u is prescribed on the boundary, that is $\partial u(t, \mathbf{x})/\partial \mathbf{n} = g_2(\mathbf{x})$, $\mathbf{x} \in \partial\Omega$ (g_2 is a given function).
- Robin boundary condition. The linear combination of u and its normal derivative is prescribed on the boundary, that is $\partial u(t, \mathbf{x})/\partial \mathbf{n} + \alpha u(t, \mathbf{x}) = g_3(\mathbf{x})$, $\mathbf{x} \in \partial\Omega$ (g_3 is a given function).

If D is constant (say one) in (1.1.1), then the equation can be written in the form

$$\nabla^2 u + f = 0 \ \text{ in } \Omega, \tag{1.1.2}$$

where ∇^2 is the so-called Laplace operator,

$$\nabla^2 u = \frac{\partial^2 u}{\partial x_1^2} + \ldots + \frac{\partial^2 u}{\partial x_d^2}$$

(d is the dimension of Ω). Equation (1.1.2) is called *Poisson equation* and in case of $f = 0$ *Laplace equation*.

In this lecture, we solve the Poisson equation with Dirichlet boundary condition on one and two-dimensional domains using the finite difference method.

1.2 One-dimensional Poisson equation with Dirichlet boundary condition

1.2.1 Setting up the problem

The one-dimensional Poisson equation with Dirichlet boundary condition (so-called one-dimensional boundary problem) is a second order ordinary differential equation in the form

$$u'' + f = 0, \quad x \in (0, 1);$$
$$u(0) = \mu_1, \ u(1) = \mu_2.$$

We suppose that the solution $u(x)$ is sufficiently smooth on $(0, 1)$ and continuous on $[0, 1]$. Of course, the above problem can be solved exactly, integrating twice both sides of the equation and choosing the two constants of integration according the

boundary conditions. Nevertheless, we solve this problem numerically with the finite difference method, in order to understand the relations between the global error and the local truncation error, and to observe how stability connects these two types of errors.

1.2.2 The finite difference solution

Let us divide the interval $[0,1]$ into $n+1$ equal parts with the points

$$0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1,$$

where $x_i = ih$ $(i = 0, \ldots, n+1)$ with $h = 1/(n+1)$. Moreover, let us introduce the notation u_i for the approximation of $\hat{u}_i = u(x_i)$ $(i = 0, \ldots, n+1)$. Naturally, these values depend on h, albeit this is not indicated in the notation, furthermore $u_0 = \mu_1$ and $u_{n+1} = \mu_2$. We have n unknown values to compute: $u_1, \ldots, u_n$. If we replace u'' by the centered difference approximations at the points x_i $(i = 1, \ldots, n)$, then we obtain the system of linear algebraic equations

$$\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + f(x_i) = 0, \quad (i = 1, \ldots, n),$$

which, multiplying each equation by -1, can be written in matrix form

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h, \tag{1.2.3}$$

where $\mathbf{u}_h = [u_1, \ldots, u_n]^\top \in \mathbb{R}^n$,

$$\mathbf{f}_h = [f(x_1) + \mu_1/h^2, f(x_2), \ldots, f(x_{n-1}), f(x_n) + \mu_2/h^2]^\top$$

and $\mathbf{A}_h$ is the tridiagonal matrix

$$\mathbf{A}_h = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \ldots & & \\ -1 & 2 & -1 & 0 & \ldots & \\ 0 & -1 & 2 & -1 & 0 & \ldots \\ & \ddots & \ddots & \ddots & \ddots & \ddots \\ & \ldots & 0 & -1 & 2 & -1 \\ & & \ldots & 0 & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{1.2.4}$$

The subscript h indicates the dependence on h. As we will see later system (1.2.3) can be solved uniquely, thus the finite difference solution do exists.

1.2.3 Stability of the method

Let us define the so-called global error as $\mathbf{z}_h = \mathbf{u}_h - \hat{\mathbf{u}}_h$. Our goal is to prove the convergence of the method, that is to show that the magnitude of the global error

goes to zero if h tends to zero ($\|\mathbf{z}_h\| \to 0$ as $h \to 0$). The magnitude of the global error can be measured, for instance, in maximum norm,

$$\|\mathbf{z}_h\|_\infty = \max_{i=1,\ldots,n} \{|z_i|\},$$

or alternatively in 2-norm (Euclidean-norm),

$$\|\mathbf{z}_h\|_2 = \sqrt{h(z_1^2 + \ldots + z_n^2)}.$$

It is known that the centered difference approximation of the second order derivative of a sufficiently smooth function gives a second order accurate approximation, so we might hope that the global error also has the order of $O(h^2)$. This is not trivial, because it is not the second derivative of the function $u(x)$ that is approximated by finite differences, but the approximating values $u_1, \ldots, u_n$ are computed using the exact values of the second derivative of $u(x)$.

Let us define the local truncation error as the vector

$$\mathbf{w}_h = \mathbf{f}_h - \mathbf{A}_h \hat{\mathbf{u}}_h, \tag{1.2.5}$$

where we substitute the vector of the exact solutions into the reordered equation (1.2.3). Of course $\hat{\mathbf{u}}$ does not generally satisfy the equation $\mathbf{A}_h \hat{\mathbf{u}}_h = \mathbf{f}_h$ exactly, the discrepancy between the two sides is the local truncation error. Combining equations (1.2.3) and (1.2.5) we obtain the system of linear equations

$$\mathbf{A}_h(\mathbf{u}_h - \hat{\mathbf{u}}_h) = \mathbf{A}_h \mathbf{z}_h = \mathbf{w}_h$$

for the global error. If $\mathbf{A}_h^{-1}$ exists, then applying the properties of induced matrix norms we have the estimation

$$\|\mathbf{z}_h\| = \|\mathbf{A}_h^{-1}\mathbf{w}_h\| \leq \|\mathbf{A}_h^{-1}\|\|\mathbf{w}_h\|.$$

We call the numerical method *consistent* if $\|\mathbf{w}_h\| \to 0$ as $h \to 0$. The numerical method is said to be *stable* if $\mathbf{A}_h^{-1}$ exists for all sufficiently small step size h, and there is a constant C, independent of h, such that

$$\|\mathbf{A}_h^{-1}\| \leq C$$

for all sufficiently small h. It is easy to see that

$$\text{stability} + \text{consistency} \Longrightarrow \text{convergence},$$

that is stability and consistency imply convergence, which is stated in general form in the so-called Lax theorem (P. Lax, 1953).

1.2.4 Consistency of the numerical scheme

Consistency is usually the easy part to check in verifying the convergence of a numerical scheme. Albeit we do not know the true solution $u(x)$, assuming the solution to be sufficiently smooth, we obtain with Taylor series expansion that

$$w_i = \frac{1}{12}h^2 u''''(x_i) + O(h^4) = O(h^2), \quad (i = 1, \ldots, n).$$

This shows that $\|\mathbf{w}_h\| \to 0$ ($h \to 0$), that is the method is consistent.

1.2.5 Stability in maximum norm

In this section the stability of the numerical method will be proven introducing the notion of M-matrices.

DEFINITION 1.2.1. A square matrix $\mathbf{A}$ is said to be an M-matrix, if each off-diagonal element of the matrix is non-positive ($a_{ij} \leq 0$) and there exists a positive vector $\mathbf{r} > \mathbf{0}$ such that $\mathbf{Ar} > \mathbf{0}$.

Theorem 1.2.2 *(See Stoyan-Takó, Numerikus módszerek I., ELTE-TypoTEXBudapest, 1993) If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is an M-matrix, then*
(P1) $\mathbf{A}$ is regular ($\exists \mathbf{A}^{-1}$),
(P2) $\mathbf{A}^{-1}$ is non-negative ($\mathbf{A}^{-1} \geq \mathbf{0}$, so-called monotone matrix),
(P3) the estimation

$$\|\mathbf{A}^{-1}\|_\infty \leq \frac{\|\mathbf{r}\|_\infty}{\min_{i=1,\ldots,n}(\mathbf{Ar})_i}$$

is valid.

The maximum norm of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is defined as

$$\|\mathbf{A}\|_\infty = \max_{i=1,\ldots,n} \{|a_{i,1}| + |a_{i,2}| + \ldots + |a_{i,n}|\}.$$

Theorem 1.2.3 *The matrix $\mathbf{A}_h$ in (1.2.4) is an M-matrix.*

PROOF. It can be seen that the off-diagonal elements of the matrix defined in (1.2.4) are non-positive, moreover let us define the vector $\mathbf{r} \in \mathbb{R}^n$ as $r_i = p(x_i)$ $(i = 1, \ldots, n)$, where $p(x) = 1/4 - (x - 1/2)^2$. It is trivial that $r_i > 0$, that is $\mathbf{r}$ is a positive vector, and considering the relation

$$-\frac{p(x_{i-1}) - 2p(x_i) + p(x_{i+1})}{h^2} = -p''(x_i) = 2$$

($p(x)$ is a quadratic polynomial) we have $\mathbf{A}_h\mathbf{r} = 2\mathbf{e} > \mathbf{0}$, where $\mathbf{e} = [1, \ldots, 1]^\top \in \mathbb{R}^n$. This completes the proof. ■

Because $\mathbf{A}_h$ is an M-matrix, $\mathbf{A}_h$ is regular, which imply that the numerical solution always exists. Moreover the estimation

$$\|\mathbf{A}_h^{-1}\|_\infty \leq \frac{1}{8}$$

holds. Thus the numerical method is stable in maximum norm. The magnitude of the global error is equal to the magnitude of the truncation error, that is $\|\mathbf{z}_h\|_\infty = O(h^2)$, thus the finite difference method is convergent.

1.2.4. EXERCISE. Prove the stability of the one-dimensional boundary problem in the Euclidean norm. (Hint: Because $\mathbf{A}_h^{-1}$ is symmetric, its Euclidean norm is

equal to its spectral radius. The eigenvalues of matrix $\mathbf{A}_h$ can be written in the form $\lambda_i = (2/h^2)(1 - \cos(i\pi h))$, $(i = 1, \ldots, n)$). □

1.2.5. Exercise. Solve problem

$$u'' + f = 0, \quad x \in (0, 1);$$

$$u'(0) = \mu_3, \ u(1) = \mu_2$$

with the finite difference method. Approximate the zero derivative on the left-hand side with the one-sided expression

$$\frac{u_1 - u_0}{h} = 0$$

or, alternatively, apply equations

$$\frac{u_{-1} - 2u_0 + u_1}{h^2} + f(x_0) = 0, \quad \frac{u_1 - u_{-1}}{2h} = \mu_3.$$

Prove the convergence of the above two methods. Compare the order of the two global error. □

1.3 Two-dimensional Poisson equation with Dirichlet boundary condition

1.3.1 Setting up the problem

After the discussion of the one-dimensional model problem, we apply the finite difference method for the two dimensional Poisson equation with Dirichlet boundary condition

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f = 0, \quad (x, y) \in \Omega = (0, a) \times (0, b);$$

$$u(x, y) = g(x, y), \quad (x, y) \in \partial\Omega.$$

We suppose that the solution $u(x, y)$ is sufficiently smooth on Ω and continuous on $\bar{\Omega}$.

1.3.2 The finite difference solution

Let us divide the edges $[0, a]$ and $[0, b]$ of Ω into $n_1 + 1$ and $n_2 + 1$ equal parts, respectively, and define the step sizes $h_1 = 1/(n_1 + 1)$ and $h_2 = 1/(n_2 + 1)$. Let us denote the interior mesh points by $P_1, \ldots, P_N$, and the points on the boundary by $P_{N+1}, \ldots, P_{N+N_\partial}$. We also define $\bar{N} = N + N_\partial$. Naturally, $N = n_1 n_2$, $N_\partial = 2(n_1 + n_2) + 4$ and $\bar{N} = (n_1 + 2)(n_2 + 2)$. Denoting the approximation of the true

solution $\hat{u}_i = u(P_i)$ at a grid point P_i by u_i, and replacing the second derivatives with centered finite differences we obtain the system of linear equations

$$\frac{u_{i-x} - 2u_i + u_{i+x}}{h_1^2} + \frac{u_{i-y} - 2u_i + u_{i+y}}{h_2^2} + f(P_i) = 0, \quad i = 1, \ldots, N. \tag{1.3.6}$$

Here we used the notation P_{i+x} for the next grid point in positive x-direction and the notation P_{i-x} for the one in negative x-direction. P_{i-y} and P_{i+y} are defined similarly (see the five-point stencil in Figure 1.3.1). Multiplying both sides by (-1)

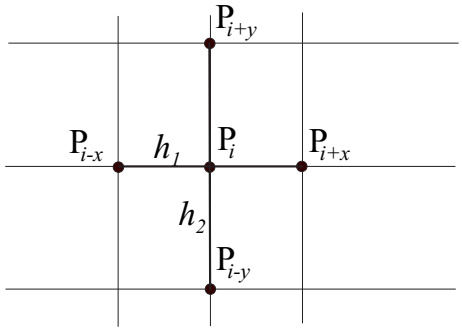

Figure 1.3.1: *The five-point stencil.*

the system can be written in matrix form as

$$\tilde{\mathbf{A}}_h \tilde{\mathbf{u}}_h = \tilde{f}_h,$$

where

$$\tilde{\mathbf{f}}_h = [f(P_1), \ldots, f(P_N)]^\top,$$
$$\tilde{\mathbf{u}}_h = [u_1, \ldots, u_{\bar{N}}]^\top$$

and matrix $\tilde{\mathbf{A}}_h$ is an $N \times \bar{N}$ sparse matrix with the i-th row elements $-1/h_1^2, -1/h_2^2$ and $2/h_1^2 + 2/h_2^2$ in the columns corresponding to the points P_{i-x} and P_{i+x}, P_{i-y} and P_{i+y}, and P_i, respectively. Considering the relations $u_i = g(P_i)$ $(i = N + 1, \ldots, N + N_\partial)$, we have only N unknown values: $u_1, \ldots, u_N$. Splitting the matrix $\tilde{\mathbf{A}}_h$ in the form $\tilde{\mathbf{A}}_h = [\mathbf{A}_h | \mathbf{A}_\partial]$ $(\mathbf{A}_h \in \mathbb{R}^{N \times N}, \mathbf{A}_\partial \in \mathbb{R}^{N \times N_\partial})$ and the vector $\tilde{\mathbf{u}}_h$ as $\tilde{\mathbf{u}}_h = [\mathbf{u}_h^\top | \mathbf{g}_h^\top]^\top$ $(\mathbf{u}_h = [u_1, \ldots, u_N]^\top \in \mathbb{R}^N$, $\mathbf{g}_h = [g(P_{N+1}), \ldots, g(P_{\bar{N}})]^\top \in \mathbb{R}^{\partial N})$ we have the equation

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h := \tilde{\mathbf{f}}_h - \mathbf{A}_\partial \mathbf{g}_h.$$

This equation has similar form like the one in the one-dimensional case. The elements of the vector $\mathbf{f}_h$ and the matrix $\mathbf{A}_h$ are known. In order to obtain the numerical solution we have to compute the vector $\mathbf{u}_h$, which consists of the approximating values for the true solution at the mesh points. Convergence means in this case that $\|\mathbf{z}_h\| \to 0$ as $h := \max\{h_1, h_2\} \to 0$. Our goal is to prove the convergence, which is implied by consistency and stability.

1.3.3 Consistency

We have to calculate the local truncation error $\mathbf{w}_h = \mathbf{f}_h - \mathbf{A}_h\hat{\mathbf{u}}_h$. Assuming the solution to be sufficiently smooth, we obtain with Taylor series expansion that

$$w_i = \frac{h_1^2}{12}\frac{\partial^4 u}{\partial x^4}(P_i) + \frac{h_2^2}{12}\frac{\partial^4 u}{\partial y^4}(P_i) + O(h_1^4) + O(h_2^4) = O(h^2), \quad (i = 1, \ldots, N).$$

This shows that $\|\mathbf{w}_h\|_\infty \to 0$ $(h \to 0)$, that is the method is consistent.

1.3.4 Stability in maximum norm

We will show that $\mathbf{A}_h$ is an M-matrix. This will show the existence and uniqueness of the finite difference solution and imply the stability.

The off-diagonal elements of $\mathbf{A}_h$ are trivially non-positive and the suitable positive vector can be constructed easily. Let us define the vector $\mathbf{r} = [p(P_1), \ldots, p(P_N)]^\top$ with the quadratic function

$$p(x, y) = \frac{a^2 + b^2}{4} - \left(x - \frac{a}{2}\right)^2 - \left(y - \frac{b}{2}\right)^2. \tag{1.3.7}$$

Clearly $\mathbf{r} > 0$, and because $p(x, y)$ is quadratic we have

$$-\frac{p(P_{i-x}) - 2p(P_i) + p(P_{i+x})}{h_1^2} - \frac{p(P_{i-y}) - 2p(P_i) + p(P_{i+y})}{h_2^2} =$$

$$= -\left(\frac{\partial^2 p}{\partial x^2}(P_i) + \frac{\partial^2 p}{\partial y^2}(P_i)\right) = 4, \quad i = 1, \ldots, N.$$

Thus $\mathbf{A}_h\mathbf{r} = 4\mathbf{e} > \mathbf{0}$, that is $\mathbf{A}_h$ is an M-matrix and the estimation

$$\|\mathbf{A}_h^{-1}\|_\infty \le \frac{a^2 + b^2}{16}$$

is true. For the global error we have $\|\mathbf{z}_h\|_\infty \le \frac{a^2+b^2}{16}\|\mathbf{w}_h\|_\infty = O(h^2)$, that is the method is convergent with second order.

1.3.1. Exercise. Let us apply the two types of ordering (rowwise and chessboard) depicted in Figure 1.3.2 in numbering the inner mesh points of the mesh. Sketch the structure of the matrix $\mathbf{A}_h$ in both cases. □

1.3.2. Exercise. Prove the stability of the finite difference solution of the two-dimensional Poisson equation using the Euclidean norm. For simplicity suppose that $h_1 = h_2 = h$ and $n_1 = n_2 = n$. □

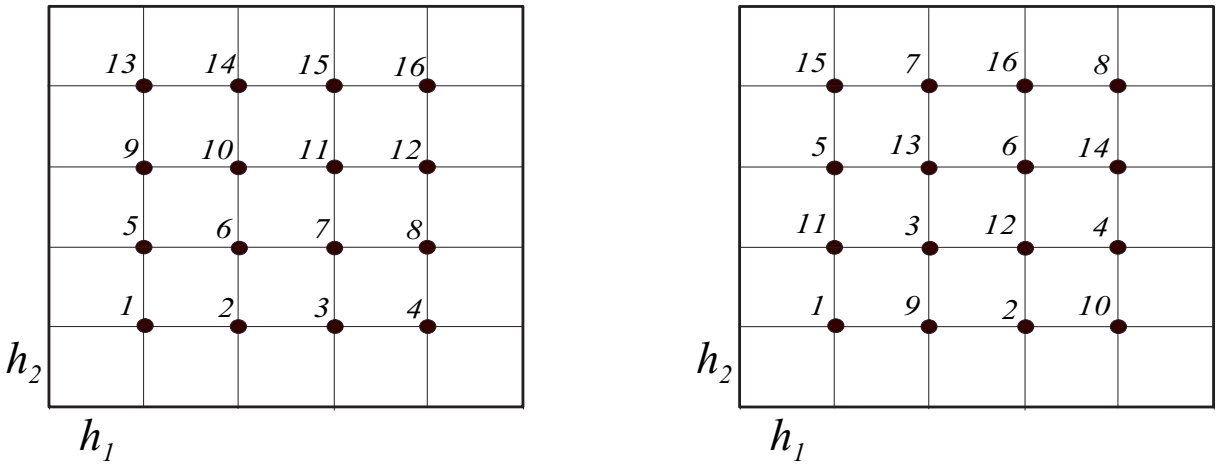

Figure 1.3.2: *The rowwise and chessboard ordering of the grid points.*

1.4 Solution of the linear system

As we discussed in the previous sections the finite difference solution of elliptic equations results in systems of linear algebraic equations. In one dimension with uniform step-size, the matrix $\mathbf{A}_h$ is a uniformly tridiagonal sparse symmetric matrix, while in two-dimension, albeit $\mathbf{A}_h$ is symmetric, the elements of the matrix cannot be clustered adjacent to the main-diagonal of the matrix. The best thing we can do is the so-called rowwise ordering (see Exercise 1.3.1). What kind of methods can be used for the solution of the linear systems? There are two different ways of the solution: direct methods and iterative methods. Direct methods produces an exact solution (in exact arithmetic), while iterative methods result in a vector sequence which converges to the solution of the system. Direct methods have advantages solving systems with dense matrices (with few zeros). The best known direct method is the so-called Gaussian elimination. Here we have to store the whole coefficient matrix and the number of operations is $O(N^3)$ (if $\mathbf{A}_h \in \mathbb{R}^{N\times N}$). Applying Gaussian elimination for systems coming from finite difference methods, there are possibilities to simplify the method taking the advantage of the special structure of the coefficient matrix achieving a number of operations $O(N)$ (so-called Thomas-algorithm). Iterative methods are generally applied for sparse systems. The crucial question here the speed of the convergence of the vector sequence to the solution of the system. The most typical iterative methods, beyond the classical Jacobi and Gauss-Seidel iterations, are successive overrelaxation, conjugate gradient method and multigrid methods. For more details consult the book *Stoyan-Takó, Numerikus módszerek I., ELTE-TypoTEXBudapest, 1993.*

Chapter 2

Finite Difference Solution of Hyperbolic Partial Differential Equations

2.1 Setting up the problem

In this chapter we will discuss finite difference methods for hyperbolic partial differential equations in one space dimension. Two types of equations are investigated. The first one is the so-called advection equation or one-way wave equation, which is second order, and has the form

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = f, \quad x \in (0,1) \tag{2.1.1}$$

$$u(0,x) = u_0(x),$$

$$u(t,0) = \mu_1(t), \text{ if } a \geq 0, \quad \text{or} \quad u(t,1) = \mu_2(t), \text{ if } a \leq 0.$$

The second one is the so-called second order linear wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2} + f, \quad x \in (0,1), \quad a > 0 \tag{2.1.2}$$

$$u(0,x) = u_0(x),$$

$$u(t,0) = \mu_1(t), \; u(t,1) = \mu_2(t).$$

The solutions of the above equations are waves, $u = u(t,x)$ gives the amplitude of the wave at time instant t and the spatial coordinate x, moreover $f = f(t,x)$ describes the density of outer forces. Because the stability of a difference scheme is usually independent of the source term f, we consider the above equations supposing that f is equal to zero.

If $f = 0$, then (2.1.1) has a solution $u(t,x) = u_0(x - at)$ (we suppose that u_0 is sufficiently smooth). As time evolves, the initial data simply propagates unchanged

to the right ($a > 0$) or to the left ($a < 0$). This is why we need boundary condition only at one of the ends of the interval. Equation (2.1.2) can be written as

$$\left(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial t} - a\frac{\partial u}{\partial x}\right) = 0,$$

which shows that the solution is a superposition of two waves

$$u(t,x) = u_{01}(x - at) + u_{02}(x + at).$$

The first one propagates to the right, and the second one to the left.

For the sake of further simplification we apply so-called periodic boundary condition which is defined as $u(0,t) = u(1,t)$. Beside this choice we do not need to discretize the boundary conditions but the differential equation.

2.2 Numerical solution of the one-way wave equation

We solve problem

$$\begin{aligned} \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad x \in (0,1) \\ u(0,x) = u_0(x), \\ u(t,0) = u(t,1) \end{aligned} \tag{2.2.3}$$

by the finite difference method. As usual, we start with dividing the interval $[0,1]$ into equal parts, the number of subintervals is chosen – for simplicity – to be $n-1$ ($h = 1/n$), furthermore, we choose a time step $\tau > 0$. Let us denote the approximation of $\hat{u}_j^k := u(k\tau, jh)$ by $u_j^k = u(k\tau, jh)$ ($j = 1,\ldots,n$; $k = 0,1,\ldots$). Naturally, $u_j^0 = u_0(jh)$ ($j = 1,\ldots,n$).

2.2.1 First try

Discretization

A natural discretization of the one-way wave equation is

$$\frac{u_j^{k+1} - u_j^k}{\tau} + a\frac{u_{j+1}^k - u_{j-1}^k}{2h} = 0,$$

where a forward and a centered finite difference is used in time and spatial coordinates, respectively. Using the notation $q = a\tau/h$, we obtain the explicit form

$$u_j^{k+1} = u_j^k - (q/2)(u_{j+1}^k - u_{j-1}^k). \tag{2.2.4}$$

These equations can be written also in an explicit matrix form as

$$\mathbf{u}_h^{k+1} = \mathbf{A}_h \mathbf{u}_h^k,$$

where $\mathbf{u}_h^k = [u_1^k, \dots, u_n^k]^\top$ and $\mathbf{A}_h$ is a skew-symmetric matrix

$$\mathbf{A}_h = \begin{bmatrix} 1 & -q/2 & 0 & \dots & 0 & q/2 \\ q/2 & 1 & -q/2 & 0 & \dots & \\ 0 & q/2 & 1 & -q/2 & 0 & \dots \\ & \ddots & \ddots & \ddots & \ddots & \ddots \\ & \dots & 0 & q/2 & 1 & -q/2 \\ -q/2 & 0 & \dots & 0 & q/2 & 1 \end{bmatrix}. \tag{2.2.5}$$

Definition of convergence

We say that the numerical method is convergent if $|u_j^k - \hat{u}_j^k| \to 0$ as $\tau, h \to 0$, where the point $(t, x) = (k\tau, jh)$ is fixed ($k, j \to \infty$). That is the method is convergent if the numerical solution at a fixed point tends to the true solution if the mesh is refined. For simplicity we suppose that τ and h tend to zero keeping the $q = a\tau/h$ to be constant.

Introducing the global error at the k-th time level as $\mathbf{z}_h^k = \mathbf{u}_h^k - \hat{\mathbf{u}}_h^k$, and the local approximation error as

$$\mathbf{w}_h^k = \frac{1}{\tau}\left(\hat{\mathbf{u}}_h^{k+1} - \mathbf{A}_h \hat{\mathbf{u}}_h^k\right),$$

we have

$$\mathbf{z}_h^{k+1} = \mathbf{A}_h \mathbf{z}_h^k - \tau \mathbf{w}_h^k.$$

Thus the global error at the k-th time level can be written in the form

$$\mathbf{z}_h^k = \mathbf{A}_h^k \mathbf{z}_h^0 - \tau \sum_{l=1}^{k} \mathbf{A}_h^{k-l} \mathbf{w}_h^{l-1},$$

from which we obtain

$$\|\mathbf{z}_h^k\| \le \|\mathbf{A}_h^k\| \|\mathbf{z}_h^0\| + \tau \sum_{l=1}^{k} \|\mathbf{A}_h^{k-l}\| \|\mathbf{w}_h^{l-1}\|.$$

If we apply the choice $u_j^0 = u_0(jh)$, then the first term vanishes. The convergence of the method can be guaranteed by supposing that the powers of the matrix $\mathbf{A}_h$ are uniformly bounded. Uniform boundedness means that $\|\mathbf{A}_h^k\| \le C$ for all sufficiently small τ, h and for each natural number k where C is independent of k and h. This is the so-called Lax-Richtmyer stability condition. The consistency and stability imply the convergence of the method. Indeed, using that the local truncation error is $O(\tau + h^2)$, we obtain that $\|\mathbf{z}_h^k\| \le \tau k \cdot C \cdot O(\tau + h^2) = t \cdot C \cdot O(\tau + h^2)$ (t is fixed), which shows the convergence. One can also see that the order of the global error is equal to the one of the local truncation error. We remark that the Lax-Richtmyer stability condition is satisfied if $\|\mathbf{A}_h\| \le 1$, since $\|\mathbf{A}_h^k\| \le \|\mathbf{A}_h\|^k \le 1$, but this condition is not necessary. If we have only $\|\mathbf{A}_h\| \le 1 + c_1\tau$ (c_1 is constant) then we also have uniform boundedness $\|\mathbf{A}_h^k\| \le (1 + c_1\tau)^k \le e^{c_1\tau k} = e^{c_1 t} = \text{constant}$.

2.2.1. Exercise. Prove that the local truncation error of the above method is $O(\tau + h^2)$. □

Convergence in Euclidean norm

We would like to show the uniform boundedness of $\mathbf{A}_h$ in the Euclidean norm. It is easy to see that the eigenvalues of the matrix are

$$\lambda_l = 1 - q\mathtt{i}\sin(2\pi lh) = 1 - \frac{\tau a}{h}\mathtt{i}\sin(2\pi lh), \quad \mathtt{i} = \sqrt{-1}, \quad (l = 1, \ldots, n)$$

This can be shown based on the consideration that matrices appearing in any standard finite difference methods have the same eigenvectors in the form $\mathbf{u}_h = [u_1, \ldots, u_n]^\top$, where $u_j = \exp(\mathtt{i}2\pi ljh)$ with $l = 1, \ldots, n$. Inserting $u_j = \exp(\mathtt{i}2\pi ljh\xi)$ into (2.2.4) we have that

$$\mathbf{A}_h\mathbf{u}_h = (1 - q\mathtt{i}\sin(2\pi lh))\mathbf{u}_h,$$

that is the l-th eigenvalue of $\mathbf{A}_h$ is λ_l, indeed. The matrix $\mathbf{A}_h$ is the sum of the unit matrix $\mathbf{I}$ (all eigenvalues are equal to one) and a skew-symmetric matrix, which eigenvalues are purely imaginary and have the form $(-\tau a/h)\mathtt{i}\sin(2\pi lh)$. Thus the square of the Euclidean norm of $\mathbf{A}_h^k$ is

$$\|\mathbf{A}_h^k\|_2^2 = \left(\varrho\left(\mathbf{A}_h^\top \cdot \mathbf{A}_h\right)\right)^k = \left(\max_l\{1 + \frac{\tau^2 a^2}{h^2}\sin^2(2\pi lh)\}\right)^k \approx (1 + q^2)^k \to \infty$$

($\varrho(.)$ denotes the spectral radius of the matrix), which shows that this method is not stable, thus it is not convergent either.

The eigenvalue λ_l is also called *amplification factor* or *growth factor* (so-called von Neumann analysis). The growth factor simply expresses the growth in the "amplitude" of the eigenvector between to time levels. It is usually denoted by $g(\theta)$ with $\theta = 2\pi lh$, and because $|g(\theta)|$ gives the Euclidean norm of $\mathbf{A}_h$ we have the same stability conditions like for the Lax-Richtmyer stability. Thus stability can be guaranteed satisfying the relation $|g(\theta)| \leq 1$ or the weaker condition $|g(\theta)| \leq 1 + c_1\tau$. If $|g(\theta)| = c_2 = \text{constant} > 1$, then the method is unstable.

The growth factor can be written for the above investigated finite difference method as $g(\theta) = (1 - q\mathtt{i}\sin\theta)$, thus $|g(\theta)|^2 = 1 + q^2\sin^2\theta > 1$, that is the method is unstable.

Our first try to solve the one-way wave equation with the finite difference method resulted in a non-convergent method, thus this method in not applicable in practice.

2.2.2 Lax-Friedrichs method

Modify the finite difference scheme used in the previous section changing u_j^k with the average of u_{j-1}^k and u_{j+1}^k. So we obtain the scheme

$$u_j^{k+1} = \frac{1}{2}(u_{j-1}^k + u_{j+1}^k) - (q/2)(u_{j+1}^k - u_{j-1}^k),$$

which is called Lax-Friedrichs scheme. The local truncation error of the scheme is $O(\tau + h)$, and the growth factor is

$$g(\theta) = \cos\theta - q\mathtt{i}\sin\theta$$

Therefore,

$$|g(\theta)|^2 = \cos^2\theta + q^2\sin^2\theta = 1 + (q^2-1)\sin^2\theta$$

and we can conclude that $|g(\theta)| \leq 1$ if $q^2 - 1 \leq 0$, which implies that

$$\frac{|a|\tau}{h} \leq 1,$$

which is called Courant-Friedrichs-Lewy (shortly CFL) condition. Albeit the method is convergent for values $|q| \leq 1$, the method is not popular because of its slow convergence.

2.2.2. Exercise. Prove that the local truncation error of the Lax-Friedrichs method is $O(\tau + h)$ and verify the expression for the growth factor! □

2.2.3 The upwind scheme

The upwind scheme for the one-way wave equation is defined as

$$\frac{u_j^{k+1} - u_j^k}{\tau} = \begin{cases} \frac{-a}{h}(u_j^k - u_{j-1}^k), & \text{if } a \geq 0 \\ \frac{-a}{h}(u_{j+1}^k - u_j^k), & \text{if } a \leq 0 \end{cases}$$

The growth factor is for the case $a \geq 0$

$$g(\theta) = 1 - q + q\cos\theta - \mathtt{i}q\sin\theta.$$

Therefore,

$$|g(\theta)|^2 = 1 - 4(1-q)q\sin^2(\theta/2)$$

and we can conclude the CFL condition $q = a\tau/h \leq 1$. Thus the upwind scheme is convergent. This scheme has the order $O(\tau + h)$ again. In the next sections we discuss some higher order schemes.

2.2.3. Exercise. Prove that the local truncation error of the upwind scheme is $O(\tau + h)$ and verify the expression for the growth factor! □

2.2.4 The leapfrog scheme

The leapfrog scheme for the one-way wave equation is defined as

$$\frac{u_j^{k+1} - u_j^k}{2\tau} + a\frac{u_{j+1}^k - u_{j-1}^k}{2h} = 0.$$

We can notice that thee time levels are involved, and we need the u_j^1 values to get started with the method.