3

The finite difference method

"Read Euler: he is our master in everything."
Pierre-Simon Laplace (1749-1827)

"Euler: The unsurpassed master of analytic invention."
Richard Courant (1888-1972)

The finite difference approximations for derivatives are one of the simplest and of the oldest methods to solve differential equations. It was already known by L. Euler (1707-1783) ca. 1768, in one dimension of space and was probably extended to dimension two by C. Runge (1856-1927) ca. 1908. The advent of finite difference techniques in numerical applications began in the early 1950s and their development was stimulated by the emergence of computers that offered a convenient framework for dealing with complex problems of science and technology. Theoretical results have been obtained during the last five decades regarding the accuracy, stability and convergence of the finite difference method for partial differential equations.

In this chapter, we will study several numerical finite difference schemes for elliptic, parabolic and hyperbolic one-dimensional boundary value problems. By means of these toy model examples and generalizations, we will define the notions of *consistency*, *stability* and *convergence* of a numerical scheme. Section 3.1 presents the general principles underlying the finite difference approximations. In Section 3.2, we adress the numerical resolution of a generic one-dimensional second-order elliptic boundary value problem. Section 3.3 focus on the finite difference method for solving the parabolic heat equation introduced in Chapter 1. In Section 3.4, we show how to generalize these results to a one-dimensional hyperbolic advection equation. Other partial differential models, like the wave equation, can also be discretized with finite difference schemes as shown in Section 3.5. The extension of the finite difference method to a higher dimensional space dimension is presented in the

Section 3.6. And to conclude this chapter, numerical experiment results are proposed in Section 3.7 to emphasize the main features and aspects related to the implementation of the finite difference method on computers.

3.1 Finite difference approximations

This chapter initiates our study of (time-dependent) partial differential equations, whose solutions vary both in time and in space. The reader may wonder why we need numerical methods, having already found analytical formulas for solving such equations in Chapter 1. Actually, there are a lot of interesting problems that cannot be solved by analytical methods, like the Fourier method. There are also several reasons for this: nonlinear problems, variable coefficients, integrals difficult to evaluate analytically, approximation based on a truncation of infinite series required, etc. And, for the problems that could be solved analytically, we are already dependent of a numerical procedure to plot the solution (cf. Chapter **??**). These observations clearly motivate the analysis of numerical methods in a more general setting.

Although there are several different numerical methods available for solving elliptic and parabolic equations, we focus here on *finite difference schemes.* This is motivated by the fact that they are very simple to understand and they are easy to generalize to more complex boundary value problems, and last but not least, they are easy to implement on a computer.

3.1.1 General principle

The principle of finite difference methods is close to the numerical schemes used to solve ordinary differential equations (cf. Appendix **??**). The simplest approach to solving PDEs numerically is to set up a regular *grid* in space and time and to compute approximate solutions at the space or time points of this grid. The essential point is the *discretization.* It consists in approximating the differential operator by replacing the derivatives in the equation using *differential quotients.* The error between the numerical solution and the exact solution is determined by the error that is commited by going from a differential operator to a difference operator. This error is called the *discretization error* or *truncation error.* The term truncation error reflects the fact that a finite part of a Taylor series is used in the approximation. For the sake of simplicity, we will consider the one-dimensional case only.

The main concept behind any finite difference scheme is related to the definition of the derivative of a smooth function u at a point $x \in \mathbb{R}$:

$$u'(x) = \lim_{h \to 0} \frac{u(x+h) - u(x)}{h}\,, \tag{3.1}$$

and to the fact that when h tends to 0 (without vanishing), the quotient on the right-hand side provides a “good” approximation of the derivative.

In other words, h should be sufficiently small to get a good approximation. It remains to indicate what exactly is a good approximation, in what sense. Actually, the approximation is good when the error commited in this approximation (*i.e.* when replacing the derivative by the differential quotient) tends towards zero when h tends to zero. If the function u is sufficiently smooth in the neighborhood of x, it is possible to quantify this error using a Taylor expansion.

3.1.2 Taylor series

Suppose the function u is C^2 continuous in the neighborhood of x. For any $h > 0$ we have:

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x+h_1) \tag{3.2}$$

where h_1 is a number between 0 and h: $x+h_1 \in]x, x+h[$. For most problems, it is convenient to retain only the first two terms of the previous expression:

$$u(x+h) = u(x) + hu'(x) + O(h^2) \tag{3.3}$$

where the term $O(h^2)$ indicates that the error of the approximation is proportional to h^2. From the Equation (3.2), we deduce that there exists a constant $C > 0$ such that, for $h > 0$ sufficienty small, we have:

$$\left| \frac{u(x+h) - u(x)}{h} - u'(x) \right| \leq C\,h\,, \qquad C = \sup_{y \in [x, x+h_0]} \frac{|u''(y)|}{2}\,, \tag{3.4}$$

for $h \leq h_0$, $h_0 > 0$ given. The error commited by replacing the derivative $u'(x)$ by the differential quotient is of *order* h. The approximation of u' at point x is said to be *consistent* at the first order. The approximation (3.3) is known as the *forward difference* approximant of u'. More generally, we define an approximation at order p of the first derivative:

Definition 3.1. *The approximation of the derivative u' at point x is of order p ($p > 0$) if there exists a constant $C > 0$, independent of h, such that the error between the derivative and its approximation is bounded by Ch^p,* i.e., *is exactly $O(h^p)$.*

Likewise, we can define the first order *backward difference* approximation of u' at point x as:

$$u(x-h) = u(x) - hu'(x) + O(h^2)\,. \tag{3.5}$$

Obviously, other approximations of u' can be considered. In order to improve the accuracy of the approximation, we define a consistent approximation, called the *central difference* approximation[1], by taking the points $x - h$ and

[1] Sometimes called the *symmetric difference* quotient.

$x+h$ into account. Suppose that the function u is three times differentiable in the vicinity of x, *i.e.*, $u \in C^3(\bar{\Omega})$ where $\Omega =]x-h, x+h[$, then we write:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u^{(3)}(\xi^+) \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u^{(3)}(\xi^-) \end{aligned} \tag{3.6}$$

where $\xi^+ \in]x, x+h[$ and $\xi^- \in]x-h, x[$. By subtracting these two expressions we obtain, thanks to the intermediate value theorem:

$$\frac{u(x+h)-u(x-h)}{2h} = u'(x) + \frac{h^2}{6}u^{(3)}(\xi)\,, \tag{3.7}$$

where ξ is a point of $]x-h, x+h[$. Hence, for every $h \in]0, h_0[$, we have the following bound on the approximation error:

$$\left|\frac{u(x+h)-u(x-h)}{2h} - u'(x)\right| \leq C\,h^2\,, \quad C = \sup_{y\in[x-h_0,x+h_0]} \frac{|u^{(3)(y)}|}{6}\,. \tag{3.8}$$

This defines a second-order consistent approximation to u' at point x.

Remark 3.1. The order of the approximation is related to the *regularity* of the function u. If u is C^2 continuous, then the approximation is consistent at the order one only.

Exercise 3.1. 1. Find a second-order approximation of $u'(x)$ involving the points x, $x+h$ and $x+2h$.
2. Find an approximation of $u'(x)$ at the order 4 based on the points x and $x+jh$, $1 \leq j \leq 4$.

3.1.3 Approximation of the second derivative

Lemma 3.1. *Suppose u is a C^4 continuous function on an interval $[x-h_0, x+h_0]$, $h_0 > 0$. Then, there exists a constant $C > 0$ such that for every $h \in]0, h_0[$ we have:*

$$\left|\frac{u(x+h)-2u(x)+u(x-h)}{h^2} - u''(x)\right| \leq C\,h^2\,. \tag{3.9}$$

The differential quotient $\frac{1}{h^2}(u(x+h)-2u(x)+u(x-h))$ is a consistent second-order approximation of the second derivative u'' of u at point x.

Proof. We use Taylor expansions up to the fourth order to achieve the result:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi^+) \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi^-) \end{aligned} \tag{3.10}$$

where $\xi^+ \in]x, x+h[$ and $\xi^- \in]x-h, x[$. Like previously, the intermediate value theorem allows us to write:

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{12} u^{(4)}(\xi)\,,$$

for $\xi \in]x-h, x+h[$. Hence, we deduce the relation (3.9) with the constant

$$C = \sup_{y \in [x-h_0, x+h_0]} \frac{|u^{(4)}(y)|}{12}\,.$$

□

Remark 3.2. Likewise, the error estimate (3.9) depends on the regularity of the function u. If u is C^3 continuous, then the error is of order h only.

3.2 Finite difference method for a 1d elliptic problem

We consider a bounded domain $\Omega =]0,1[\subset \mathbb{R}$ and $u : \bar{\Omega} \to \mathbb{R}$ solving the non-homogeneous Dirichlet problem:

$$(\mathcal{D}) \begin{cases} -u''(x) + c(x)u(x) = f(x)\,, & x \in \Omega\,, \\ u(0) = \alpha\,, \;\; u(1) = \beta\,, & \alpha, \beta \in \mathbb{R}\,, \end{cases} \tag{3.11}$$

where c and f are two given functions, defined on $\bar{\Omega}$, $c \in C^0(\bar{\Omega}, \mathbb{R}_+)$. Such system can be used to represent the diffusion-reaction phenomenon related to a chemical species in a substrate.

3.2.1 Variational theory and approximation

Since Chapter 2, we know that if $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$, then the solution u to the boundary value problem $(\mathcal{D})$ exists. Furthermore, if $c \equiv 0$, we have the explicit formulation of u as:

$$u(x) = \int_\Omega G(x,y) f(y)\, dy + \alpha + x(\beta - \alpha)\,, \tag{3.12}$$

where $G(x,y) = x(1-y)$ if $y \geq x$ and $G(x,y) = y(1-x)$ if $y < x$. However, when $c \neq 0$ there is no explicit formula giving the solution u. Thus, we should resign to find an approximation of the solution.

The first step in deriving a finite difference approximation of the boundary value problem (3.11) is to partition the unit interval $\Omega =]0,1[$ into a finite number of subintervals. We have the fundamental concept of the finite difference approximations: the numerical solution is not defined on the whole domain Ω but at a finite number of points in Ω only.

To this end, we introduce the equidistributed grid points $(x_j)_{0\leq j\leq N+1}$ given by $x_j = jh$, where N is an integer and the spacing h between these points is given by:

$$h = \frac{1}{N+1}\,,$$

and we have then $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$. Typically, the spacing is aimed at becoming very small as the number of grid points will become very large. At the boundary of $\varOmega$, we have $x_0 = 0$ and $x_{N+1} = 1$. At each of these points, we are looking for numerical value of the solution, $u_j = u(x_j)$. Of course, we impose $u(x_0) = \alpha$ and $u(x_{N+1}) = \beta$, the Dirichlet boundary conditions and we use the differential quotient introduced in the previous section to approximate the second order derivative of the equation (3.11).

3.2.2 A finite difference scheme

The unknowns of the discrete problem are all the values $u(x_1), \ldots, u(x_N)$ and we introduce the vector $u_h \in \mathbb{R}^N$ of components u_j, for $j \in \{1, \ldots, N\}$.

Suppose functions c and f are at least such that $c \in C^0(\bar{\varOmega})$ and $f \in C^0(\bar{\varOmega})$. The problem is then to find $u_h \in \mathbb{R}^N$, such that $u_i \simeq u(x_i)$, for all $i \in \{1, \ldots, N\}$, where u is the solution of the nonhomogeneous Dirichlet problem (3.11). Introducing the approximation of the second order derivative by a differential quotient, leads to consider the following discrete problem:

$$(\mathcal{D}_h)\left\{\begin{array}{c} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + c(x_j)u_j = f(x_j)\,, \qquad j \in \{1, \ldots, N\} \\ u_0 = \alpha\,,\ \ u_{N+1} = \beta\,, \end{array}\right. \tag{3.13}$$

The problem $(\mathcal{D})$ has been discretized by a finite difference method based on a *three-points centered* scheme for the second-order derivative. This problem consists in a set of N linear equations. Two additional equations correspond to setting the boundary values using Dirichlet conditions.

The discrete problem $(\mathcal{D}_h)$ can be written in a more concise the matrix form as follows:

$$A_h u_h = b_h\,, \tag{3.14}$$

where A_h is the tridiagonal matrix defined as:

$$A_h = A_h^{(0)} + \begin{pmatrix} c(x_1) & 0 & \ldots & \ldots & 0 \\ 0 & c(x_2) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & c(x_{N-1}) & 0 \\ 0 & \ldots & \ldots & 0 & c(x_N) \end{pmatrix}, \tag{3.15}$$

with

$$A_h^{(0)} = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \qquad \text{and} \qquad b_h = \begin{pmatrix} f(x_1) + \dfrac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \dfrac{\beta}{h^2} \end{pmatrix} \tag{3.16}$$

There are at least two questions arising from this formulation:

(i) is there a (unique) solution to the system $(\mathcal{D}_h)$?
(ii) Does u_h converges toward the exact solution u ? If yes, in what sense ?

In other words, we have to determine if the matrix A_h is invertible or not. The answer is given by the following proposition.

Proposition 3.1. *Suppose $c \geq 0$. Then, the matrix A_h is symmetric positive definite, hence A_h is invertible.*

Proof. It is easy to check that the matrix A_h is symmetric. It remains to show that A_h is positive definite. Let consider a vector $v = (v_i)_{1\leq i \leq N} \in \mathbb{R}^N$. Since $c \geq 0$, we have:

$$v^t\, A_h\, v = v^t\, A_h^{(0)} v + \sum_{i=1}^{N} c(x_i) v_i^2 \;\geq\; v^t\, A_h^{(0)}\, v\,, \tag{3.17}$$

and the problem is reduced to show that $A_h^{(0)}$ is positive definite. We can notice that:

$$\begin{aligned} h^2 v^t\, A_h\, v &= v_1^2 + (v_2 - v_1)^2 + \cdots + (v_{N-1} - v_N)^2 + v_N^2 \\ &= \sum_{i=1}^{N} (v_i - v_{i-1})^2 + v_N^2 \;\geq\; 0\,, \quad \forall v = (v_1, \dots, v_N) \in \mathbb{R}^N\,, \end{aligned} \tag{3.18}$$

and thus $v^t\, A_h\, v \geq 0$. Moreover, if $v^t\, A_h\, v = 0$ then all terms $v_{i+1} - v_i = v_1 = v_N = 0$. Hence, we can conclude that all $v_i = 0$ and the result follows. □

Remark 3.3 (existence and uniqueness). Since the matrix A_h is invertible, there exists a unique solution to the problem $(\mathcal{D}_h)$. We could have achieved this result by showing that $\ker(A_h) = 0$, *i.e.*, in finite dimension, any surjective or injective linear mapping is bijective.

The second question raised the convergence issue of the finite difference method. It will be answered in the next sections.

3.2.3 Consistent scheme

The formula used in the numerical schemes result from an approximation of the equation using a Taylor expansion. The notion of *consistency* and of *accuracy* helps to understand how well a numerical scheme approximates an equation. Here, we want to know if the problem $(\mathcal{D}_h)$ is a good approximation of the problem $(\mathcal{D})$. To this end, we introduce a formal definition of the consistency that can be used for any partial differential equation defined on a bounded domain Ω and denoted as:

$$(Lu)(x) = f(x)\,, \qquad \text{for all } x \in \Omega\,, \tag{3.19}$$

where L denotes a differential operator. The notation (Lu) indicates that the equation depends on u and on its derivatives at any point $x \in \Omega$. A numerical scheme can be written, for every index j, in a more abstract form as follows:

$$(L_h u)(x_j) = f(x_j)\,, \qquad \text{for all } j \in \{1, \dots, N\}\,. \tag{3.20}$$

For example, the boundary-value problem (3.11) has the operator L defined as:

$$(Lu)(x) = -u''(x) + c(x)u(x)\,, \tag{3.21}$$

and the problem can be written in the following form:

$$\text{find } u \in C^2(\Omega) \text{ such that } (Lu)(x) = f(x)\,, \text{ for all } x \in \Omega\,. \tag{3.22}$$

Now, we define the approximation operator L_h, for all $j \in \{1, \dots, N\}$, by:

$$(L_h u)(x_j) = -\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + c(x_j)u(x_j)\,, \tag{3.23}$$

and the discrete problem (3.13) can be formulated as follows:

$$\text{find } u \text{ such that } (L_h u)(x_j) = f(x_j)\,, \qquad \text{for all } j \in \{1, \dots, N\}\,. \tag{3.24}$$

We are now able to introduce the following definition of the consistency of a numerical scheme.

Definition 3.2. *A finite difference scheme is said to be* consistent *with the partial differential equation it represents, if for any sufficiently smooth solution u of this equation, the* truncation error *of the scheme, corresponding to the vector $\varepsilon_h \in \mathbb{R}^N$ whose components are defined as:*

$$(\varepsilon_h)_j = (L_h u)(x_j) - f(x_j)\,, \qquad \textit{for all } j \in \{1, \dots, N\} \tag{3.25}$$

tends uniformly towards zero with respect to x, when h tends to zero, i.e. *if:*

$$\lim_{h \to 0} \|\varepsilon_h\|_\infty = 0\,.$$

Moreover, if there exists a constant $C > 0$, independent of u and of its derivatives, such that, for all $h \in]0, h_0]$ ($h_0 > 0$ given) we have:

$$\|\varepsilon_h\| \le C\, h^p \,,$$

with $p > 0$, then the scheme is said to be accurate at the order p with respect to the norm $\|\cdot\|$ considered.

Usually, we consider, for every vector $v \in \mathbb{R}^N$ the following norms:

$$\|v\|_\infty = \sup_{1\le i\le N} |v_i| \qquad \|v\|_1 = \sum_{i=1}^{N} |v_i| \qquad \|v\|_2 = \left(\sum_{i=1}^{N} (v_i)^2\right)^{1/2} \tag{3.26}$$

This definition states that the truncation error is defined by applying the difference operator L_h to the exact solution u. A consistent scheme means that the exact solution almost solves the discrete problem.

Lemma 3.2. *Suppose $u \in C^4(\bar{\Omega})$. Then the numerical scheme ($\mathcal{D}_h$) is consistent and second-order accurate in space for the norm $\|\cdot\|_\infty$.*

Proof. By using the fact that $-u'' + cu = f$ and if we suppose $u \in C^4(\Omega)$, we have:

$$\begin{aligned}
\varepsilon_h(x_j) &= -\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + c(x_j)u(x_j) - f(x_j) \\
&= -u(x_j) + \frac{h^2}{12}u^{(4)}(\xi_j) + c(x_j)u(x_j) - f(x_j) \\
&= \frac{h^2}{12}u^{(4)}(\xi_j)\,.
\end{aligned}$$

where each of the ξ_j belongs to the interval $]x_{j-1}, x_{j+1}[$. Hence, we have:

$$\|\varepsilon_h\|_\infty \le \frac{h^2}{12} \sup_{y\in\Omega} |u^{(4)}(y)|\,,$$

and the result follows. □

Remark 3.4. (i) Since the space dimension N is related to h by the relation $h(N+1) = 1$, we have then:

$$\|\varepsilon_h\|_1 = O(h)\,, \qquad \text{and} \qquad \|\varepsilon_h\|_2 = O(h^{3/2})\,.$$

(ii) If we denote $\bar{u}_h = (u(x_j))_{1\le j\le N} \in \mathbb{R}^N$ the vector with all its components corresponding to the exact values of the solution of $-u'' + cu = f$, then we have:

$$\varepsilon = A_h(u_h - \bar{u}_h)\,, \qquad \text{with} \quad \varepsilon \in \mathbb{R}^N\,.$$

3.2.4 The Hopf maximum principle

E. Hopf proved in 1927 that if a function satisfies a second-order partial differential inequality of a certain type in a domain of $\mathbb{R}^d$ and attains a maximum in the domain, then the function is constant. In this section, we will explain how and why this notion is important for finite difference schemes. At first, we introduce some matrix related definitions.

Definition 3.3. *Let $A = (a_{ij}) \in M_N(\mathbb{R})$. The matrix A is* positive *(or $A > 0$) if all its coefficients a_{ij} are strictly positive. It is* nonnegative *if all $a_{ij} \geq 0$. A matrix A is* monotone *if A is invertible and if A^{-1} is nonnegative, $A^{-1} \geq 0$.*

We have the following characterization of a monotone matrix.

Lemma 3.3. *A matrix $A = (a_{ij}) \in M_N(\mathbb{R})$ is* monotone *if and only if for every vector $v \in \mathbb{R}^N$, we have $Av \geq 0$ iif $v \geq 0$.*

Proof. We carry out the proof of the necessary and sufficient conditions.
(i) [*necessary condition*]. Suppose A is monotone. We consider a vector $v \in \mathbb{R}^N$ such that $Av \geq 0$, *i.e.,* $(Av)_i \geq 0$ for all $i \in \{1, \dots, N\}$. Then, since A is invertible by hypothesis, the following relation holds:

$$v = A^{-1}(Av)\,,$$

Componentwise this relation reads:

$$v_i = \sum_{j=1}^{N} (A^{-1})_{ij} (Av)_j\,,$$

with $(A^{-1})_{ij} \geq 0$. We deduce easily that $v_i \geq 0$ for all $i \in \{1, \dots, N\}$ and thus $V \geq 0$.
(ii)[*sufficient condition*]. Suppose the matrix A is such that $Av \geq 0 \Rightarrow v \geq 0$ for any vector $v \in \mathbb{R}^N$. Let show first that the matrix A is invertible. We consider a vector $w \in \ker A$, *i.e.,* such that: $Aw = 0$. Trivially, we have $Aw \geq 0 \Rightarrow w \geq 0$. Likewise, $-w \in \ker A$, hence $-w \geq 0$. We deduce that $w_i = 0$ for all $i \in \{1, \dots, N\}$, and the kernel is reduced to the null vector. Considering b_j the j^{th} column vector of the matrix A^{-1} and $e_j \geq 0$ the j^{th} vector of the canonical basis, leads to write:

$$A^{-1} e_j = b_j\,, \qquad \text{or similarly:} \quad A b_j = e_j\,.$$

We deduce that $b_j \geq 0$, for all $j \in \{1, \dots, N\}$ and thus $A^{-1} \geq 0$ and this concludes the proof. □

This result can be seen as a property of conservation of the positivity, that reveals sometimes very important in physical applications. Next, we enounce two important properties related to monotone matrices and to the maximum principle for a continuous elliptic boundary-value problem.

Proposition 3.2 (monotony and positivity). *Let $A \in M_N(\mathbb{R})$. Then A preserves the* positivity *if and only if A is* monotone*.*

Proposition 3.3 (maximum principle). *Assume that $f \in C^0(\bar{\Omega})$ is a nonnegative function. Then the* minimum *of the function u solution of the problem $(\mathcal{D})$ is attained on the boundary of the domain Ω. Furthermore, if $c \in C^0(\bar{\Omega})$ is a nonnegative function, then the corresponding solution u of $(\mathcal{D})$ is also nonnegative.*

The following result can be understood as the discrete equivalent of the maximum principle for the numerical scheme $(\mathcal{D}_h)$.

Lemma 3.4 (discrete maximum principle). *Suppose $c \in C^0(\bar{\Omega})$ is a nonnegative function,* i.e., *$c(x_j) \geq 0$, for all $j \in \{1, \dots, N\}$. Then the matrix A_h defined by the relation (3.15) is* monotone*.*

Proof. We pose $\mu = \inf\{u_j, j \in \{1, \dots, N\}\}$ and we want to show that $\mu \geq 0$. To this end, we suppose here that $\mu < 0$ and we carry out the proof by contradiction.

Let consider $J = \{j \in \{1, \dots, N\}$, such that $u_j = \mu\}$. It is obvious to check that j is not empty. We denote by k the largest element (index) in the set J. We have:

$$\begin{cases} u_k = \mu & \text{since } k \text{ is the largest in } J \\ u_k < u_{k+1} & \\ u_k \leq u_{k-1} & \end{cases} \Rightarrow 2u_k \leq u_{k+1} + u_{k-1} \Rightarrow u_{k+1} - 2u_k + u_{k-1} \geq 0\,, \tag{3.27}$$

The numerical scheme $(\mathcal{D}_h)$ gives:

$$-\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} - c_k u_k = f_k$$

and leads to the relation: $u_{k+1} - 2u_k + u_{k-1} = h^2(c_k u_k - f_k)$. The hypothesis $\mu = u_k < 0$ and $c_k \geq 0$ give $c_k u_k \leq 0$. Since $f_k \geq 0$, we conclude that $c_k u_k - f_k \leq 0$, hence $u_{k+1} - 2u_k + u_{k-1} \leq 0$, which is in contradicton with the the relation (3.27) and thus with the hypothesis $\mu < 0$.
We conclude finally that $u_k \geq 0$ for all $j \in \{1, \dots, N\}$ and thus we have the property:

$$A_h u_h \geq 0 \,\Rightarrow\, u_h \geq 0\,,$$

that characterizes a monotone matrix A_h. □

Proposition 3.4. *Suppose $c \in C^0(\bar{\Omega})$ is a nonnegative function. Then, the matrix A_h is* invertible*.*

We now turn to the fundamental notions of stability and of convergence of the numerical scheme $(\mathcal{D}_h)$.

3.2.5 Convergence and stability of the numerical scheme

Before introducing the concept of stability, we recall some basic definitions of matrix norms.

Definition 3.4. *The* norm *of a matrix $A \in M_n(\mathbb{R})$ is a real-valued function $A \mapsto \|A\|$ satisfying the following four properties:*

(i) $\|A\| > 0$ *for all* $A \in M_n(\mathbb{R})$*,* $\|A\| = 0 \Leftrightarrow A \equiv 0$.
(ii) $\|\lambda A\| = |\lambda| \, \|A\|$ *for all* $A \in M_n(\mathbb{R})$*.*
(iii) $\|A + B\| \leq \|A\| + \|B\|$*, for all* $A, B \in M_n(\mathbb{R})$*.*
(iv) $\|A\,B\| \leq \|A\| \, \|B\|$ *for all* $A, B \in M_n(\mathbb{R})$*.*

In comparison with the definition of th norm of a vector, we consider here the additional property (iv) called the *submultiplicative* property.

Definition 3.5. *Given a vector norm $\|\cdot\|$ on $\mathbb{R}^n$, the function $\|\cdot\| : M_n(\mathbb{R}) \to \mathbb{R}_+$ defined as:*

$$\|A\| = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|=1}} \|Ax\| \,, \tag{3.28}$$

is called the induced *or* subordinate *matrix norm.*

The induced norm s a matrix norm compatible with the given vector norm. It is the smallest norm among all matri xnorms that are compatible. We recall that a matrix norm is *compatible* or *mutually consistent* with the vector norm $\|x\|$ if the inequality:

$$\|Ax\| \leq \|A\| \, \|x\|$$

is valid for all $x \in \mathbb{R}^n$ and all $A \in M_n(\mathbb{R})$.

The most common matrix norms are:

a) the *total* norm: $\|A\|_G = n \max_{1 \leq i,k \leq n} |a_i k|$.

b) the *Froebenius* norm: $\|A\|_F = \left(\sum_{i,k=1}^{n} a_{ik}^2 \right)^{1/2}$.

c) the *maximum row norm*: $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{k=1}^{n} |a_{ik}|$.

d) the *maximum column sum*: $\|A\|_1 = \max_{1 \leq k \leq n} \sum_{i=1}^{n} |a_{ik}|$.

All these matrix norms are equivalent. For example, we have:

$$\frac{1}{n} \|A\|_G \leq \|A\|_p \leq \|A\|_G \leq n \|A\|_p \,, \qquad p \in \{1, \infty\} \,.$$

The matrix norm induced by the Euclidean vector norm is:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \sqrt{x^t(A^tA)x} = \sqrt{\lambda_{max}(A^tA)} = \sqrt{\rho(A^tA)}\,,$$

where $\rho(A)$ denotes the *spectral radius* of the matrix A, *i.e.,* the modulus of the eigenvalue of A having the largest modulus. This norm is also called the *spectral norm.* In the special case of a symmetric matrix A, the matrix $A^tA = A^2$ has the eigenvalues λ_i^2 satisfying:

$$\|A\|_2 = |\lambda_{max}(A)|\,.$$

Definition 3.6. *The number $\kappa(A) = \|A\|\,\|A^{-1}\|$ is called the* condition number *of the matrix A with respect to the matrix norm considered.*

The following relation holds:

$$1 \leq \|I_n\| = \|AA^{-1}\| \leq \|A\|\,\|A^{-1}\|\,.$$

Exercise 3.2. Show that the eigenvalues λ_k of the Laplacian matrix $A_h^{(0)}$ (cf. Equation (3.16)) correspond to:

$$\lambda_k = \frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(N+1)}\right)\,, \qquad \text{for } 1 \leq j \leq N\,,$$

and that the eigenvectors are given by:

$$(X_k)_j = \sin\left(\frac{jk\pi}{N+1}\right)\,, \qquad \text{for } 1 \leq j,k, \leq N\,.$$

Deduce that for N sufficiently large, the condition number of the matrix $A_h^{(0)}$ with respect to the Euclidean norm is such that:

$$\kappa_2(A_h^{(0)}) \approx \frac{4(N+1)^2}{\pi^2}\,.$$

Proposition 3.5 (stability). *Given a nonnegative function $c \in C^0(\bar{\Omega})$, the numerical scheme $(\mathcal{D}_h)$ is said to be* stable, *with respect to $\|\cdot\|_\infty$, if there exists a constant $C > 0$, independent of h, such that:*

$$\|(A_h)^{-1}\|_\infty \leq C\,, \qquad \textit{here} \quad C = \frac{1}{8}\,. \tag{3.29}$$

This inequality can be rewritten as follows:

$$\|u_h\|_\infty \leq \frac{1}{8}\|f\|_\infty\,. \tag{3.30}$$

Proof. First, we notice that:

$$A_h^{-1} - (A(0)_h)^{-1} = A_h^{-1}(A_h^{(0)} - A_h)(A_h^{(0)})^{-1}\,.$$

Since $c \geq 0$, $(A_h^{(0)} - A_h)$ is a diagonal matrix with negative coefficients, thus $A_h^{-1} \geq 0$ and $(A_h^{(0)})^{-1} \geq 0$ are monotone matrices. Hence, we deduce that:

$$A_h^{-1} \leq (A_h^{(0)})^{-1}$$

and this leads to the inequality:

$$\|A_h^{-1}\|_\infty \;\leq\; \|(A_h^{(0)})^{-1}\|_\infty \,.$$

We have simply to show that $\|(A_h^{(0)})^{-1}\|_\infty \frac{1}{8}$ to achieve te result. Since $(A_h^{(0)})^{-1}$ is a positive matrix, we have:

$$\|(A_h^{(0)})^{-1}\|_\infty = \|(A_h^{(0)})^{-1} e\|_\infty$$

where $e \in \mathbb{R}^n$ is the vector having all its components equal to one. Furthermore, $(A_h^{(0)})^{-1}$ is the solution of the linear system $A_h u_h = b_h$ in the special case where $c \equiv 0$ and for $b_h = e$, *i.e.,* it is the solution of the following homogeneous Dirichlet boundary-value problem:

$$\begin{cases} \quad -u''(x) = 1\,, & x \in \Omega \\ u(0) = u(1) = 0\,, & \end{cases} \tag{3.31}$$

The analytical solution is known: $u_0(x) = x(x-1)/2$, such that $u_0^{(4)}(x) = 0$. Involving Lemma 3.2, we know that the numerical solution coincide with the exact solution at the grid points $(x_j)_{1 \leq j \leq N}$. Hence, we have:

$$\Big((A_h^{(0)})^{-1} e\Big)_j = u(x_j)\,,$$

and we can deduce that:

$$\|(A_h^{(0)})^{-1} e\|_\infty \;\leq\; \sup_{x \in \Omega} |u(x)| = \frac{1}{8}\,.$$

This achieves to proves the estimate. □

Definition 3.7. *The* discretization error *at point x_j is the difference between the exact solution $u(x_j)$ at x_j and the j^{th} component of the solution vector corresponding to the numerical scheme:*

$$e_j = u(x_j) - u_j\,, \qquad \textit{for all } j \in \{1, \ldots, N\}\,.$$

Theorem 3.1 (convergence). *Suppose $c \in C^0(\bar{\Omega})$ is a nonnegative function and $u \in C^4(\bar{\Omega})$. Then the finite difference scheme $(\mathcal{D}_h)$ is* convergent *of* order 2 *for the norm $\|\cdot\|_\infty$ and we have:*

$$\max_{1 \leq i \leq N} |e_i| \;\leq\; \frac{1}{96} \|u^{(4)}(x)\|_\infty \, h^2\,. \tag{3.32}$$

Proof. We have $A_h(u-u_h)=\varepsilon_h$, where ε_h is the truncation error. Hence:

$$\|\tilde{u}-u_h\|_\infty \leq \|A_h^{-1}\|_\infty \|\varepsilon_h\|_\infty \leq \frac{1}{8}\frac{1}{12}\|u^{(4)}(x)\|_\infty = \frac{1}{96}\|u^{(4)}(x)\|_\infty\,.$$

where $\tilde{u}=(u(x_1),\dots,u(x_N))^t \in \mathbb{R}^N$. □

Theorem 3.2. *A* consistent *and* stable *numerical method is* convergent *and the order of convergence is at least equal to the order of consistency of the numerical scheme.*

3.2.6 Neumann boundary conditions

We consider a bounded domain $\Omega=]0,1[\subset \mathbb{R}$ and $u:\bar{\Omega}\to\mathbb{R}$ solving the following boundary-value problem:

$$(\mathcal{N})\begin{cases} -u''(x)+c(x)u(x)=f(x)\,, & x\in\Omega\,,\\ u'(0)=\alpha\,,\ u'(1)=\beta\,, & \alpha,\beta\in\mathbb{R}\,,\end{cases} \tag{3.33}$$

where c and f are two given functions defined on $\bar{\Omega}$, $c\in C^0(\bar{\Omega})$ and $f\in C^0(\bar{\Omega})$. The variational theory indicates that such a solution u exists if there exists a constant $c_0>0$ such that $c(x)\geq c_0$, for all $x\in\bar{\Omega}$.

We consider equidistributed grid points $(x_j)_{0\leq j\leq N+1}$ given by $x_j=jh$. We consider the scheme previously defined for the Dirichlet problem $(\mathcal{D})$ to compute the values of u at the internal points x_j, for all $j\in\{1,\dots,N\}$. The question that remains now is how to discretize the Neumann boundary conditions ? Several approaches can be envisaged. For instance, the first derivative of u can be discretized by a difference scheme as follows:

$$u'(0)\approx\frac{u(h)-u(0)}{h} \qquad \text{and} \qquad u'(1)\approx\frac{u(1)-u(1-h)}{h}\,, \tag{3.34}$$

thus leading to the numerical scheme:

$$(\mathcal{N}_h)\begin{cases} -\dfrac{u_{k+1}-2u_k+u_{k-1}}{h^2}+c_ku_k=f_k\\ \qquad\qquad u_0=u_1-h\alpha\\ \qquad\qquad u_{N+1}=u_N+h\beta\,,\end{cases} \tag{3.35}$$

and to solve the linear system in matrix form:

$$B_hu_h=b_h\,, \tag{3.36}$$

where the matrix B_h is defined as:

$$B_h=B_h^{(0)}+\begin{pmatrix} 0 & \dots & \dots & \dots & 0\\ \vdots & c(x_1) & & & \vdots\\ \vdots & & \ddots & & \vdots\\ \vdots & & & c(x_N) & \\ 0 & \dots & \dots & \dots & 0\end{pmatrix} \tag{3.37}$$

with

$$B_h^{(0)} = \frac{1}{h^2}\begin{pmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad b_h = \begin{pmatrix} -\dfrac{\alpha}{h} \\ f(x_1) \\ \vdots \\ \vdots \\ f(x_N) \\ \dfrac{\beta}{h} \end{pmatrix} \tag{3.38}$$

Remark 3.5. (i) Here, the unknown $u_h \in \mathbb{R}^{N+2}$ and $b_h \in M_{N+2}(\mathbb{R})$.

(ii) Under the assumption $c(x) \geq c_0 > 0$, for all $x \in \Omega$, the matrix B_h is symmetric, positive definite, thus B_h is invertible. In this case, the system $B_h u_h = b_h$ has a unique solution.

(iii)if $c \equiv 0$, the problem cannot be addressed: $\ker B_h^{(0)}$ contains the constants. The matrix B_h is monotone if $c(x) \geq 0$ and $c(x) \neq 0$ in at least one internal grid point.

The numerical scheme $(\mathcal{N}_h)$ is *consistent* and *first-order accurate* in space for the norm $\|\cdot\|_\infty$: *i.e.,* $\|\varepsilon_h\|_\infty = O(h)$. To improve the accuracy of the solution, we can use a central difference scheme for the Neumann boundary conditions:

$$u'(0) \approx \frac{u(h)-u(-h)}{2h} \qquad \text{and} \qquad u'(1) \approx \frac{u(1+h)-u(1-h)}{2h}\,, \tag{3.39}$$

however, this requires introducing additional *fictitious* unknowns corresponding to the data u_{-1} and u_{N+2}, thus increasing the size of the linear system to be solved. This leads to define the following numerical scheme:

$$(\mathcal{N}_h')\begin{cases} -\dfrac{u_{k+1}-2u_k+u_{k-1}}{h^2} + c_k u_k = f_k\,, \qquad k \in \{0,\dots,N+1\} \\ \qquad u_{-1} = u_1 - 2h\alpha \\ \qquad u_{N+2} = u_N + 2h\beta\,, \end{cases} \tag{3.40}$$

The consistency error (truncation error) of the numerical scheme $(\mathcal{N}_h')$ is then in $O(h^2)$, better than with the first-order difference approximation of the first derivative used to define the scheme $(\mathcal{N}_h)$.

3.3 Finite difference method for a 1d parabolic problem

In this section, we consider the *heat transfer* equation as an example of parabolic problem, for $a \in \mathbb{R}$:

$$\begin{cases} \dfrac{\partial u}{\partial t} - a\dfrac{\partial u^2}{\partial x^2} = 0\,, \qquad \text{for } x \in \mathbb{R}, t > 0 \\ \quad u(x,0) = u_0(x)\,, \quad x \in \mathbb{R} \end{cases} \tag{3.41}$$

3.3.1 Solutions of the heat equation

The following function $p(x,t)$, called the *Gaussian density* function, is a special solution of (3.41):

$$p(x,t) = (4\pi at)^{-1/2} \exp\left(-\frac{x^2}{4at}\right), \qquad \text{for } x \in \mathbb{R}, t > 0\,, \tag{3.42}$$

that allows to retrieve all solutions by translation and superposition. The initial values of p are difficult to exhibit:

$$p(x,t) \to 0\,, x \neq 0\,, p(x,t) \to +\infty\,, x = 0, t \to 0_+ \text{ and } \int p\,dx = 1\,, \forall t > 0\,.$$

thus $p|_{t=0} = \delta_0$, the Dirac mass at 0.Then we have:

$$u(x,t) = u_0 * p_t = \int_{\mathbb{R}} u_0(y)p(x-y,t)\,dy\,, \quad \forall x \in \mathbb{R}\,, t > 0\,,$$

$u(x,t)$ is a *regular solution* ($^\infty(x,t), t > 0$), periodic of period L in x. Moreover, if u_0 is continuous, u can be prolongated by continuity at $t = 0$ and $u|_{t=0} = u_0$ on $\mathbb{R}$. Hence, given $h > 0$:

$$\begin{aligned}
&|u(x,t) - u_0(x)| = \left|\int_{\mathbb{R}} (u_0(x-y) - u_0(x))p(y,t)\,dy\right| \\
&\leq \sup_{|y|\leq h} |u_0(x-y) - u_0(x)| \left(\int_{|y|\leq h} p(y,t)\,dy\right) + 2\sup_z |u_0(z)| \int_{|y|>h} p(y,t)\,dy \\
&\leq \sup_{|y|\leq h} |u_0(x-y) - u_0(x)| + \frac{2}{\sqrt{\pi}} \sup_z |u_0(z)| \int_{|z|\geq \frac{h}{2\sqrt{at}}} \exp\left(-\frac{z^2}{2}\right) dz\,.
\end{aligned} \tag{3.43}$$

Hence, we have:

$$\limsup_{t\to 0_+} \left(\sup_x |u(x,t) - u_0(x)\right) \leq \sup_{|x_1 - x_2|\leq h} |u_0(x_1) - u_0(x_2)|\,. \tag{3.44}$$

Proposition 3.6. *Given u_0 continuous and periodic, the function u defined by the formula (3.42) gives a solution of the heat equation (3.41), regular for $x \in \mathbb{R}, t > 0$, periodic in x, continuous on $\mathbb{R} \times [0,+\infty[$ and satisfying the initial condition $u_0(x) = u(x,0)$. Furthermore, this solution is* unique.

Remark 3.6. The heat equation has two noticeable properties:

(i) the solution is regular for $t > 0$, as the physical phenomenon of diffusion is.

(ii) the solution remains positive if u_0 is positive:

$$u(x,t) \geq 0\,, \forall(x,t) \in \mathbb{R} \times [0,+\infty[\,, \quad \text{if } u_0(x) \geq 0\,, \forall x \in \mathbb{R}\,.$$

3.3.2 Finite difference schemes

We consider the one-dimensional heat equation posed in the bounded domain $\Omega =]0,1[\subset \mathbb{R}$:

$$(\mathcal{H}) \left\{ \begin{aligned} \frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial t^2} &= 0\,, \qquad \text{for } (x,t) \in \bar{\Omega} \times \mathbb{R}_+^*\,, \\ u(x,0) &= u_0(x)\,, \qquad \text{for } x \in \Omega\,, \\ u(0,t) = u(1,t) &= 0\,, \end{aligned} \right. \tag{3.45}$$

where $a \in \mathbb{R}_+$ is a given scalar value and $u_0 \in C^0(\bar{\Omega}, \mathbb{R})$. To study this initial-boundary value problem, we proceed in a very similar way we did in the previous section for the elliptic boundary value problem. The basic idea is again to replace the derivatives involved in the equation (3.45) by finite differences.

To this end, we introduce the equidistributed grid points $(x_j)_{0 \le j \le N+1}$ given by $x_j = jh$, with $h = 1/(N+1)$, to discretize the domain $\bar{\Omega}$ in space. In addition, we consider the time $\Delta t > 0$ to discretize the domain in time, hence defining a *regular grid*, $t^n = n\Delta t$, $n \in \mathbb{N}$.

We denote u_j^n the value of a numerical solution at point (x_j, t^n) and $u(x,t)$ the exact solution of the equation (3.45). The initial data u_0 is discretized by:

$$u_j^0 = u_0(x_j) \qquad \text{for } j \in \{0, \ldots, n+1\}\,, \tag{3.46}$$

and the Dirichlet boundary conditions are translated into:

$$u_0^n = u_{N+1}^n = 0 \qquad \text{for every } n > 0\,. \tag{3.47}$$

Before defining several numerical schemes, let us recall that we consider the following approximations:

$$\begin{aligned} u_t(x,t) &= \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} + O(\Delta t) \\ u_{xx}(x,t) &= \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} + O(h^2)\,. \end{aligned} \tag{3.48}$$

These approximations naturally leads to the following *explicit* scheme:

$$(\mathcal{H}_1) \left\{ \begin{aligned} \frac{u_j^{n+1} - u_j^n}{\Delta t} - a\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} &= 0\,, \qquad j \in \{1, \ldots, N\}, n \ge 0\,, \\ u_0^{n+1} = u_{N+1}^{n+1} &= 0\,, \end{aligned} \right. \tag{3.49}$$

To initialize the scheme, se set:

$$u_j^0 = u_0(x_j)\,, \quad \text{for } j \in \{0, \ldots, N+1\}\,.$$

This scheme can be rewritten in a more convenient form as follows:

$$u_j^{n+1} = u_j^n + \frac{a\Delta t}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) , \tag{3.50}$$

and then we can observe that the values u_j^{n+1} at time t^{n+1} can be deduced from the values u_j^n at time t^n, thus justifying the label *explicit*.

An alternative consists in changing al indices n into $n+1$ in the space discretization, yielding the following *implicit* scheme:

$$(\mathcal{H}_2)\frac{u_j^{n+1} - u_j^n}{\Delta t} - a\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} = 0 , \qquad j \in \{1, \ldots, N\}, n \geq 0 , \tag{3.51}$$

This scheme is well-defined: all values u_j^{n+1} at time t^{n+1} can be computed based on the previous values u_j^n attime t^n, it requires then to invert the tridiagonal matrix $C_h \in M_N(\mathbb{R})$:

$$C_h = \begin{pmatrix} 1+2c & -c & 0 \ldots & 0 \\ -c & 1+2c & -c \ddots & \vdots \\ 0 & \ddots & \ddots \ \ddots & 0 \\ \vdots & \ddots & \ddots \ \ddots & -c \\ 0 & \ldots & 0 \ \ -c & 1-2c \end{pmatrix} \quad \text{with } c = \frac{a\Delta t}{h^2} \tag{3.52}$$

Notice that C_h is a positive definite matrix, hence is invertible.

Adding the two previous schemes and taking the mean value leads to the *Crank-Nicolson* scheme[2]:

$$(\mathcal{H}_3)\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{a}{2}\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} - \frac{a}{2}\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} = 0 , \tag{3.53}$$

More generally, we can consider a weighted average of the explicit and implicit schemes, leading to the *theta* scheme, or the θ-scheme:

$$(\mathcal{H}_4)\frac{u_j^{n+1} - u_j^n}{\Delta t} - a(1-\theta)\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} - a\theta\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} = 0 , \tag{3.54}$$

where θ denotes a positive parameter, $\theta \in [0, 1]$. Obviously, we retrieve the explicit scheme for the value $\theta = 0$, the implicit scheme for $\theta = 1$ and the Crank-Nicolson scheme for $\theta = 1/2$. The θ scheme is an implicit scheme for $\theta \neq 0$.

Other schemes have been proposed, like the *Richardson* scheme:

$$(\mathcal{H}_5)\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - a\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0 . \tag{3.55}$$

[2] J. Crank and P. Nicolson (1947), A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat-Conduction Type, *Proc. of the Cambridge Philosophical Society*, **43**, 50-67.

This scheme is an *explicit three-levels* scheme (it involves the indices $n, n+1, n-1$). By replacing u_j^n in this scheme by the mean value between u_j^{n+1} and u_j^{n-1}, we obtain the explicit *Dufort-Frankel* scheme:

$$(\mathcal{H}_6)\frac{u_j^{n+1}-u_j^{n-1}}{2\Delta t}-a\frac{u_{j+1}^n-u_j^{n+1}-u_j^{n-1}+u_{j-1}^n}{h^2}=0\,. \tag{3.56}$$

And finally, the *Gear* scheme[3] is widely used in practice:

$$(\mathcal{H}_7)\frac{3u_j^{n+1}-4u_j^n+u_j^{n-1}}{2\Delta t}-a\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{h^2}=0\,. \tag{3.57}$$

Having all these schemes at hand, raises the question of determining which one among all these schemes is more appropriate to solve the heat problem ? Numerical analysis will certainly help us to decide with respect to various criteria. One criterion is already visible: the higher computational cost of the implicit methods as compared to that of the explicit schemes. Indeed, implicit methods require the resolution of a linear system at each time step. In addition, three-levels schemes will most likely require more memory to store the information than two-levels schemes. All the schemes involve a finite number of values u_j^n. The basic structure of a numerical scheme, *i.e.,* for a given point (x_j, t^n) the neighbours that appear in the discrete scheme, is called the *stencil* of the scheme. We have considered three or four-point stencil discretizations.

Remark 3.7. The Dirichlet boundary conditions in the initial-boundary value problem $(\mathcal{H})$ can be replaced by Neumann boundary conditions or periodic boundary conditions or Fourier boundary conditions. Neumann boundary conditions can be discretized as follows:

$$\frac{\partial u}{\partial x}(0,t)=0\rightarrow\frac{u_1^n-u_0^n}{h}=0 \text{ and } \frac{\partial u}{\partial x}(1,t)=0\rightarrow\frac{u_{N+1}^n-u_N^n}{h}=0\,. \tag{3.58}$$

This allows to eliminate the values u_0^n and u_{N+1}^n and requires only computing the values $(u_j^n)_{1\le j\le N}$. Actually, this approach provides a *first-order* discretization of the Neumann conditions. A *second-order* discretization would then be:

$$\frac{u_1^n-u_{-1}^n}{2h}=0 \quad\text{and}\quad \frac{u_{N+2}^n-u_N^n}{2h}=0 \tag{3.59}$$

that requires adding two *fictitious* grid points x_{-1} and x_{N+2}.

Likewise, periodic boundary conditions can be defined as follows:

$$u(x+1,t)=u(x,t)\,,\quad \text{for every } x\in\bar{\Omega}, t\ge 0\,, \tag{3.60}$$

and can be discretized using the inequalities $u_0^n=u_{N+1}^n$, for every $n\ge 0$, or more generally by: $u_j^n=u_{N+1-j}^n$.

[3] C.W. Gear (1968), The automatic integration of stiff ordinary differential equations, *Proc. IFIP Congress*, Ljubjana, Yugoslavia, 81-85.

Remark 3.8. In the case of an existing right-hand side $f(x,t)$ in the initial-boundary value problem $(\mathcal{H})$, the numerical schemes can be adjusted to take this information into account. This is achieved by introducing a consistent approximation of $f(x,t)$ at the grid points (x_j, t^n). For example, the explicit scheme $(\mathcal{H}_1)$ becomes:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - a\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = f(x_j, t^n)\,, \tag{3.61}$$

3.3.3 Consistent scheme

In this section, we consider the explicit scheme $(\mathcal{H}_1)$ to analyze the notions of *consistency*, of *accuracy* and of *truncation error.*

Definition 3.8 (consistency). *The numerical scheme $(\mathcal{H}_1)$ is said to be* consistent *with the partial differential operator* $\dfrac{\partial}{\partial t} - a\dfrac{\partial^2}{\partial x^2}$, *if for any sufficiently smooth solution $u = u(x,t)$ of this equation, the difference:*

$$\left(\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} - a\frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2}\right) \\ -(\partial_t u - a\partial_{xx}u)(x,t) = (\varepsilon_h u)(x,t) \tag{3.62}$$

tends toward 0 when both h and Δt tend to 0 independently. Moreover, if there exists a constant $C > 0$, independent of u and of its derivatives, such that $(\varepsilon_h u)(x,t)$ tends toward 0 as $O(h^p + \Delta t^q)$ when h and Δt tend to 0:

$$\|\varepsilon_h u\|_\infty \leq C(h^p + \Delta t^q)\,, \qquad p > 0, q > 0\,, \tag{3.63}$$

the scheme is said to be accurate *at the order p in space and the order q in time.*

Proposition 3.7. *The explicit scheme $(\mathcal{H}_1)$ is* consistent, first-order accurate *in time and* second-order accurate *in space for the approximation of the heat equation $u_t - au_{xx} = 0$, if $u \in C^{2,4}(\bar{\Omega}, [0,T])$.*

Proof. The result is based on Taylor expansions. We evaluate:

$$\frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\Delta t} - u_t(x_j, t^n)\,.$$

Since u is supposed C^2 continuous in time, there exists $\tau \in [t^n, t^{n+1}]$ such that:

$$u(x_j, t^{n+1}) = u(x_j, t^n) + \Delta t\, u_t(x_j, t^n) + \frac{\Delta t^2}{2} u_{tt}(x_j, \tau)\,,$$

and thus:

$$\frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\Delta t} - u_t(x_j, t^n) = \frac{\Delta t^2}{2} u_{tt}(x_j, \tau)\,.$$

Similarly, we evaluate:

$$\frac{u(x_{j+1},t^n)-u(x_j,t^n)+u(x_{j-1},t^n)}{h^2}=u_{xx}(x_j,t^n)\,.$$

Since u is supposed C^4 continuous in space, there exists $\xi_+\in[x_j,x_{j+1}]$ such that:

$$\begin{aligned}u(x_{j+1},t^n)=u(x_j,t^n)+hu_x(x_j,t^n)+\frac{h^2}{2}u_{xx}(x_j,t^n)&\\+\frac{h^3}{6}u_{xxx}(x_j,t^n)+\frac{h^4}{24}u_{xxxx}(\xi_+,t^n)&,\end{aligned}$$

and likewise, there exists $\xi_-\in[x_{j-1},x_j]$. Hence, we obtain:

$$\begin{aligned}\frac{u(x_{j+1},t^n)-2u(x_j,t^n)+u(x_{j-1},t^n)}{h^2}-u_{xx}(x_j,t^n)=&\\\frac{h^2}{24}\left(u_{xxxx}(\xi_+,t^n)+u_{xxxx}(\xi_-,t^n)\right)&.\end{aligned}$$

Now, we introduce the heat equation $u_t(x_j,t^n)-au_{xx}u(x_j,t^n)=0$ to obtain:

$$(\varepsilon_h u)_j^n=\frac{\Delta t}{2}u_{tt}(x_j,\tau)-a\frac{h^2}{24}(u_{xxxx}(\xi_+,t^n)+u_{xxxx}(\xi_-,t^n))\,,\qquad(3.64)$$

and the desired result is achieved by introducing the constant:

$$C=\frac{1}{2}\max\left(\max_{\bar{\Omega}\times[0,T]}|u_{tt}|,\frac{a}{6}\max_{\bar{\Omega}\times[0,T]}|u_{xxxx}|\right)\qquad(3.65)$$

that yields to the estimate:

$$|(\varepsilon_h u)_j^n|\le C(h^2+\Delta t)\,,\quad\forall j\in\{1,\dots,N\},\forall n\in\{0,\dots,M\}\,.\qquad(3.66)$$

If we denote $e^n(u)$ the error vector at time t^n: $e^n(u)=(e_j^n(u))_{1\le j\le N}$ with

$$e_j^n(u)=u_j^n-u(x_j,t^n)\,,\qquad\forall j\in\{1,\dots,N\}\,,n\in\{0,\dots,M\}\,,$$

and assuming that

$$a\frac{\Delta t}{h^2}\le\frac{1}{2}\,,\qquad(3.67)$$

then there exists $C\in\mathbb{R}$ such that:

$$\|e^n(u)\|_\infty\le C(h^2+\Delta t)\,,\qquad\forall n\ge 0\,.\qquad(3.68)$$

The table 3.1 reports the order of the truncation errors for the numerical schemes.

Remark 3.9. The condition $a\Delta t/h^2 \leq 1/2$ that relates the time step Δt to the space step h is sometimes called the *Courant-Friedrichs-Lewy condition* (the CFL condition[4] in short) by analogy with the transport equations. It is indeed a very restrictive condition that requires the time step to be such that:

$$\Delta t \leq \frac{h^2}{2a}$$

.

3.3.4 Maximum principle

We provide now an interesting property of the numerical schemes $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$ similar to the maximum principle.

Lemma 3.5 (maximum principle). *Let $\sigma \in \mathbb{R}_+^*$ and $u_0 \in C^0(\bar{\Omega})$. Then the solution u of the heat problem (3.41) endowed with the Dirichlet boundary conditions verifies for all $(x,t) \in \bar{\Omega} \times \mathbb{R}_+$:*

$$\min\left(\min_{y\in\bar{\Omega}} u_0(y), 0\right) \leq u(x,t) \leq \max\left(\max_{y\in\bar{\Omega}} u_0(y), 0\right) . \tag{3.69}$$

Proof. It is based on the maximum principle for the differences $u - \max_{y\in\bar{\Omega}} u_0(y)$ and $u - \min_{y\in\bar{\Omega}} u_0(y)$ (see Remark 3.6). □

Lemma 3.6 (discrete maximum principle). *Let (u_j^n) be the numerical solution obtained using the explicit scheme $(\mathcal{H}_1)$. Under the assumption:*

$$a\frac{\Delta t}{h^2} \leq \frac{1}{2},$$

Numerical scheme	Truncation error
implicit $(\mathcal{H}_2)$	$O(h^2 + \Delta t)$
Crank-Nicolson $(\mathcal{H}_3)$	$O(h^2 + \Delta t^2)$
θ scheme $(\mathcal{H}4)$, $\theta \neq \frac{1}{2}$	$O(h^2 + \Delta t)$
Richardson $(\mathcal{H}_5)$	$O(h^2 + \Delta t^2)$
Dufort-Frankel $(\mathcal{H}_6)$	$O(h^2 + \frac{\Delta t^2}{h^2})$
Gear $(\mathcal{H}_7)$	$O(h^2 + \Delta t^2)$

Table 3.1. *Order of the truncation errors for various numerical schemes.*

[4] R. Courant, K.O. Friedrichs and H. Lewy (1928), Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.*, **100**, 32-74.

we have:

$$\min\left(\min_{0\leq j\leq N+1} u_j^0, 0\right) \leq u_j^n \leq \max\left(\max_{0\leq j\leq N+1} u_j^0, 0\right), \quad \forall n \geq 0. \qquad (3.70)$$

Proof. We pose $\lambda = \dfrac{a\Delta t}{h^2}$ and we write the explicit scheme as follows:

$$u_j^{n+1} = \lambda u_{j+1}^n + (1-2\lambda)u_j^n + \lambda u_{j-1}^n .$$

Assuming that $\lambda \leq 1/2$, the coefficients λ, $1-2\lambda$ of the convex combination are all positie (or null) and since the sum $2\lambda + 1 - 2\lambda = 1$, we deduce:

$$\min(u_{j-1}^n, u_j^n, u_{j+1}^n) \leq u_j^{n+1} \leq \max(u_{j-1}^n, u_j^n, u_{j+1}^n),$$

and the result follows using a recurrence relation. □.

Remark 3.10. It is interesting to notice that the *implicit* scheme $(\mathcal{H}_2)$ preserves the maximum principle without the restriction on the time step, *i.e.,* without the CFL condition.

3.3.5 Convergence of the scheme

At this stage, we are ready to enounce a convergence result for the explicit scheme $(\mathcal{H}_1)$.

Theorem 3.3 (convergence of the explicit scheme $(\mathcal{H}_1)$). *Suppose the exact solution of problem (refheatpde) is C^4 continuous in space and is C^2 continuous in time in $\bar{\Omega} \times \mathbb{R}_+$. Suppose also that $a \in \mathbb{R}_+^*$ and $N \in \mathbb{N}^*$. Let (u_j^n) be the numerical solution obtained using the explicit scheme $(\mathcal{H}_1)$, with $u_j^0 = u_0(x_j)$, $0 \leq j \leq N+1$. We suppose also that:*

$$\lambda = \frac{a\Delta t}{h^2} \leq \frac{1}{2}.$$

Then, for all $T > 0$, there exists a constant $C_T > 0$ (independent of h, depending on u and T only) such that:

$$\sup_{\substack{0\leq j\leq N+1\\ 0\leq t^n\leq T}} |u_j^n - u(x_j,^n)| \leq C_T\,(h^2 + \Delta t). \qquad (3.71)$$

Proof. We can write (cf. Proposition 3.7):

$$\frac{u(x_j,t^{n+1}) - u(x_j,t^n)}{\Delta t} - a\frac{u(x_{j+1},t^n) - 2u(x_j,t^n) + u(x_{j-1},t^n)}{h^2}$$
$$= \frac{\Delta t}{2}u_{tt}(x_j, t^n + \mu_j^n) - a\frac{h^2}{12}u_{xxxx}(x_j + \nu_j^n, t^n),$$

with $0 \leq \mu_j^n \leq \Delta t$ and $-h \leq \nu_j^n \leq h$. We call ε_j^h the right-hand side and, $T > 0$ being fixed, we denote:

$$K = \sup_{\substack{0 \leq x \leq 1 \\ 0 \leq t \leq T}} \left(|u_t t(x,t)| + |u_{xxxx}(x,t)| \right).$$

When $n\Delta t \leq T$, we obtain the following estimate:

$$|\varepsilon_j^n| \leq K\frac{\Delta t}{2} + \frac{aKh^2}{12}. \tag{3.72}$$

We introduce the numerical error $e_j^n = u_j^n - u(x_j, t^n)$. By difference, we obtain:

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} - a\frac{e_{j+1}^n - e_j^n + e_{j-1}^n}{h^2} = -\varepsilon_j^n,$$

hence:

$$e_j^{n+1} = \lambda e_{j-1}^n + (1-2\lambda)e_j^n + \lambda e_{j-1}^n - \Delta t \varepsilon_j^n, \qquad \text{with } \lambda = a\frac{\Delta t}{h^2}.$$

Denoting $E^n = \sup |e_j^n|$, $j = 0, \ldots, N+1$, and since $1 - 2\lambda \geq 0$ and $\lambda > 0$, we have:

$$|e_j^{n+1}| \leq E^n + \Delta t \left(\frac{K\Delta t}{2} + a\frac{Kh^2}{12} \right),$$

and thus:

$$E^{n+1} \leq E^n + \Delta t\frac{K}{2}\left(1 + \frac{a}{6}\right)(h^2 + \Delta t).$$

Noticing that $E^0 = 0$ ($u_j^0 = u_0(x_j), \forall j$), we can deduce easily that:

$$E^{n+1} \leq n\Delta t\frac{K}{2}\left(1 + \frac{a}{6}\right)(h^2 + \Delta t), \qquad \text{hence } C_t = \frac{KT}{2}\left(1 + \frac{a}{6}\right).$$

The numerical scheme $(\mathcal{H}_1)$ is *convergent* when $a\Delta t/h^2 \leq 1/2$ since then:

$$\lim_{h, \Delta t \to 0} \left(\sup_{\substack{0 \leq j \leq N+1 \\ 0 \leq t^n \leq T}} |u_j^n - u(x_j, t^n)| \right) = 0. \tag{3.73}$$

Remark 3.11. The estimate (3.71) on the error reflects that the explicit scheme is accurate at the order one in time and at the order two in space. This estimate can only be obtained on a bounded interval $[0, T]$ in time and not for $t \in \mathbb{R}_+$, although it does not not depend on the number N of grid points in Ω.

3.3.6 Stability and convergence

The proof of Theorem (3.3) shows that the convergence of the explicit scheme $(\mathcal{H}_1)$ relies on the discrete maximum principle (cf. Proposition 3.7). Here, the maximum principle can be seen as a *stability property* that ensures that, under the assumption $\lambda \leq 1/2$, the u_j^n values remain bounded uniformly, for every j and every n values. The proof of the convergence of the numerical scheme involves three ingredients: the *linearity* of the numerical scheme, the *consistency* and the *stability* of the scheme.

Denoting $u^n = u(^n_j)_{1\leq j\leq N} \in \mathbb{R}^N$ the solution at time $t^n = n\Delta t$, we introduce the following norms:

$$\|u^n\|_1 = \sum_{j=1}^N h|u_j^n|\,, \quad \|u_j^n\|_2 = \left(\sum_{j=1}^N h|u_j^n|^2\right)^{1/2}\,, \quad \|u_j^n\|_\infty = \sup_{1\leq j\leq N} |u_j^n|\,.$$

These norms are the discrete equivalents of the $L^1(\Omega)$, $L^2(\Omega)$ and $L^\infty(\Omega)$ norms.

Definition 3.9 (stability). *A numerical scheme is said to be* stable *with respect to a norm $\|\cdot\|$ if and only if there exists a constant $K > 0$, independent of h and of Δt, such that:*

$$\|u^n\| \leq K\,\|u^0\|\,, \qquad \forall n \in \mathbb{N}\,, \tag{3.74}$$

whatever the initial data u^0 could be.

This definition and Lemma 3.6 show that the explicit scheme $(\mathcal{H}_1)$ is *stable* in norm $L^\infty(\Omega)$ under the condition: $a\Delta t/h^2 \leq 1/2$.

Remark 3.12. The implicit numerical scheme $(\mathcal{H}_2)$ is *unconditionnably stable* in $L^\infty(\Omega)$ norm.

Theorem 3.4 (Lax's theorem). *Given u the exact solution of the initial-boundary value problem (3.41) and (u_j^n) the numerical solution obtained using a consistent two-levels linear scheme, with $u_j^0 = u_0(x_j)$, for all j. If the solution u is sufficiently smooth and if the numerical scheme is stable with respect to a norm $\|\cdot\|$, then the numerical method* converges *as:*

$$\forall T > 0\,, \qquad \lim_{h,\Delta t\to 0}\left(\sup_{t^n\leq T}\|e^n\|\right) = 0\,, \tag{3.75}$$

where $e_j^n = u_j^n - u(x_j, t^n)$ and $e^n = (e_j^n)_{q\leq j\leq N}$.
Furthermore, if the numerical scheme is consistent at the order p in space and at the order q in time, then for any $T > 0$, there exists a constant $C_t > 0$ such that:

$$\sup_{t^n\leq T}\|e^n\| \leq C_T(h^p + \Delta t^q)\,. \tag{3.76}$$

Proof. A linear two-levels scheme can be written as follows:

$$u^{n+1} = A_h u^n \,,$$

where $A_h \in M_N(\mathbb{R})$ is the iteration matrix. We denote by $\tilde{u}^n = (\tilde{u}^n_j)_{1\leq j\leq N}$ with $\tilde{u}^n_j = u(x_j, t^n)$. Since the numerical scheme is consistent, there exists a vector ε^n such that:

$$\tilde{u}^{n+1} = A_h \tilde{u}^n + \Delta t\, \varepsilon^n \,, \qquad \text{with } \lim_{h,\Delta t\to 0} \|\varepsilon^n\| = 0 \,,$$

and the convergence of ε^n is uniform for all times $0 \leq t^n \leq T$. Moreover, we have:

$$\varepsilon^{n+1} = A_h e^n - \Delta t\, \varepsilon^n \,, \qquad \text{with } e^n_j = u^n_j - u(x_j, t^n)$$

since $\|\varepsilon^n\| \leq C(h^p + \Delta t^q)$. By recurrence, we obtain:

$$e^n = A^n_h e^0 - \Delta t \sum_{k=1}^{N} A_h^{n-k} \varepsilon^{k-1}$$

and the stability condition indicates that $\|u^n\| = \|A^n_h u^0\| \leq K\|u^0\|$, for any initial data, *i.e.,* $\|A^n_h\| \leq K$, K being independent of n. Since $e^0 = 0$, we obtain finally:

$$\|e^n\| \,\leq\, \Delta t \sum_{k=1}^{N} \|A_h^{n-k}\| \, \|\varepsilon^{k-1}\| \,\leq\, n\, \Delta t\, KC(h^p + \Delta t^q) \,.$$

and the results follows, with $C_T = TKC$. □

Remark 3.13. Actually, the original Lax's *equivalence* theorem states that a *linear* and *consistent* finite difference scheme *converges* if and only if it is *stable.* From the numerical point of view, it is usually sufficient to establish the stability of the scheme to ensure its convergence. We will see in the examples that a unstable scheme cannot converges.

Proposition 3.8. *Suppose $a\Delta t/h^2 \leq 1/2$, then the numerical scheme ($\mathcal{H}_1$) converges in L^2-norm: if the solution u of the problem 3.45 is sufficiently smooth, then for every $T > 0$:*

$$\lim_{h,\Delta t\to 0} \left(\sup_{t^n \leq T} \|e^n\|_2 \right) = 0 \,, \tag{3.77}$$

where $e^n_j = u^n_j - u(x_j, t^n)$ and $e^n = (e^n_j)_{1\leq j\leq N}$,

Proof. The scheme ($\mathcal{H}_1$) can be written as: $u^{n+1} = A_h u^n$ with

$$A_h = \begin{pmatrix} 1-2\lambda & \lambda & 0 & \dots & 0 \\ \lambda & 1-2\lambda & \lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \lambda \\ 0 & \dots & 0 & \lambda & 1-2\lambda \end{pmatrix} \qquad \text{with } \lambda = a\frac{\Delta t}{h^2}\,.$$

Hence, we have $A_h = I - a\Delta t A_N$, where $I \in M_N(\mathbb{R})$ is the identity matrix and the matrix A_N is defined as follows:

$$A_N = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

The eigenvalues of A_h, denoted $(\lambda_k)_{1\le k\le N}$ are given by:

$$\lambda_k = 1 - 4\lambda \sin^2\left(\frac{k\pi h}{2}\right), \qquad 1 \le k \le N\,,$$

and are such that $\lambda_N \le \lambda_k \le \lambda_1 \le 1$ for $1 \le k \le N$ as $h = 1/(N+1)$, or

$$\lambda_N = 1 - 4\lambda\cos^2\left(\frac{\pi h}{2}\right) \le \lambda_k \le 1\,, \qquad 1 \le k \le N\,,$$

and thus $\lambda_N \ge -1$ and $\rho(A_h) \le 1$ if $\lambda \le 1/2$. We recall that $\|A_h\|_2 = \rho(A_h)$ for a symmetric matrix A_h, hence we write:

$$\|u^n\|_2 \le \|A_h^n\|_2\|u^0\|_2 = \rho(A_h^n)\|u^0\|_2 = \rho(A_h)^n\|u^0\|_2 \le \|u^0\|_2\,,$$

this proves that under the condition $\lambda \le 1/2$, the explicit scheme $(\mathcal{H}_1)$ is sable in L^2-norm and it is thus convergent with respect to this norm, thanks to the Lax's theorem. □

Lemma 3.7. *Consider the θ-scheme ($\mathcal{H}_4$), for $\theta \in [0,1]$.*

(i) If $1/2 \le \theta \le 1$, the θ-scheme is unconditionnally stable *in L^2-norm (and convergent).*

(ii)If $0 \le \theta < 1/2$, the θ-scheme is stable *in L^2-norm (and convergent for this norm) under the condition:*

$$a\frac{\Delta t}{h^2} \le \frac{1}{2(1-2\theta)}\,. \tag{3.78}$$

Lemma 3.8. *The Dufort-Frankel scheme ($\mathcal{H}_6$) and the Gear scheme ($\mathcal{H}_7$) are* unconditionnally stable *and* convergent *in L^2-norm.*
The Richardson scheme ($\mathcal{H}_5$) is unconditionnably unstable *in L^2-norm.*

In Section 3.7, we will provide various examples of the solutions obtained by the schemes described here to illustrate the concepts developed in this section.

3.4 Finite difference method for a 1d hyperbolic problem

The linear *advection* equation $u_t + cu_x = 0$ ($c \in \mathbb{R}_+$ given) is undoubtably one of the simplest of all partial differential models, albeit computing an accurate numerical solution on a fixed grid is still a topic of active research in numerical analysis. In this section, we will introduce the finite difference method to solve this linear equation. To this end, we consider the Cauchy problem:

$$\begin{cases} u_t + cu_x = 0\,, & (x,t) \in \mathbb{R} \times \mathbb{R}_+ \\ u(x,0) = u_0(x)\,, & x \in \mathbb{R}\,. \end{cases} \tag{3.79}$$

We suppose the initial data u_0 of class C^1 and periodic on $\mathbb{R}$, of period $L > 0$.

Remark 3.14. The periodicity hypothesis is not strictly required for the analysis of the problem. It allows us to reduce the numerical resolution of the problem (3.79) to a bounded interval. It is thus required from the practical point of view.

We recall that the exact solution of the Cauchy problem is given by:

$$u(x,t) = u_0(x - ct)\,, \tag{3.80}$$

that shows that for $t > 0$ given, $u(x,t)$ is a periodic function in the space variable x, of period L. Hence, we restrict the analysis to $[0,L] \times \mathbb{R}_+$.

We could have considered a boundary-value problem posed in a bounded domain $\Omega = [0,L] \subset \mathbb{R}$:

$$\begin{cases} u_t + cu_x = 0\,, & (x,t) \in \Omega \times \mathbb{R}_+ \\ u(x,0) = u_0(x)\,, & x \in \mathbb{R}\,, \\ u(0,t) = g(t)\,, & t \in \mathbb{R}_+ \end{cases} \tag{3.81}$$

however, the Cauchy problem (3.79) with the periodicity condition is more general, as we don't need to care for the discretization of the boundary conditions.

3.4.1 Method of characteristics

As already mentioned, if $u_0 \in C^1(\mathbb{R})$, the Cauchy problem (3.79) has a unique solution $u \in C^1(\mathbb{R} \times \mathbb{R}_+)$ given by:

$$u(x,t) = u_0(x - ct)\,, \qquad c(x,t) = c \;\in \mathbb{R}, \forall x, t\,, \tag{3.82}$$

this can be easily verified by applying the change of variables $(x,t) \mapsto (\eta,\xi)$ defined by:

$$\eta = x + ct\,, \qquad \xi = x - ct\,, \qquad \text{if } c \neq 0\,.$$

The *characteristics* are curves in the half-plane $\mathbb{R} \times \mathbb{R}_+$ defined as follows: for a given $x_0 \in \mathbb{R}$, consider the ordinary differential equation:

$$\begin{cases} \dfrac{dx(t)}{dt} = c(x(t),t)\,, \\ x(0) = x_0\,. \end{cases} \tag{3.83}$$

The solution $x = x(t)$ of this ODE problem defines a curve $\{(x(t),t), t \geq 0\}$ starting at $(x_0, 0)$ at time $= 0$.

Now, we want to consider a function $u \in C^1(\mathbb{R} \times \mathbb{R}_+)$ along the characteristics, *i.e.,* we like to study the evolution of $u(x(t),t)$. By differentiating u with respect to t we get:

$$\frac{d}{dt}u(x(t),t) = u_t + c\,u_x\frac{dx}{dt} = (u_t + c\,u_x)(x(t),t)\,. \tag{3.84}$$

Hence, we can state the following property.

Proposition 3.9. *If $u \in C^1$ verifies the advection equation $u_t + c\,u_x = 0$, then u is constant along the characteristic curves of equation $x - ct = k$.*

Note that in this linear problem, the characteristics cannot cross so long as the given function $c(x,t)$ is Lipschitz-continuous in x and continuous in t. When the function c is a constant, the characteristics are the parallel straight lines $x - ct = k$ and the solution is then:

$$u(x,t) = u_0(x - ct)\,.$$

On the other hand, if $c = c(u)$ is a function only of u, the characteristics are also straight lines, u is constant along each, although they are no longer parallel. The exact solution of the nonlinear problem can the be written as:

$$u(x,t) = u_0(x - c(u(x,t))t)\,,$$

until the time when the characteristics intersect or cross each other.

In the applications, we shall consider quasi-linear *systems of conservation laws* of the form:

$$u_t(x,t) + A(u(x,t))u_x(x,t) = 0\,, \tag{3.85}$$

where $u : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n$ is a vector of unknown functions and $A(u) \in M_n(\mathbb{R})$ depends continuously on u. The *hyperbolicity* of the system is related to the fact that we assume that A is diagonalizable and has real eigenvalues. The system is *stricly* hyperbolic if and only if it is hyperbolic and all its eigenvalues are distinct, for all u.

Suppose we denote by Λ the diagonal matrix of eigenvalues, and by $P = P(u)$ the matrix of left eigenvectors so that $\Lambda = P^{-1}\,A\,P$. Then, a function u is solution of the system if and only if:

$$P^{-1}u_t + P^{-1}\,\Lambda\,PP^{-1}u_x = 0\,, \tag{3.86}$$

or similarly:

$$v_t + \Lambda v_x = 0\,,$$

where v is a vector of *Riemann invariants* $v = v(u)$ such that $v_t = Pu_t$ and $v_x = Pu_x$, or $A : v = P^{-1}u$. This is a direct generalization of the scalar case. Each component of the diagonal matrix Λ will depend on all the components of v, the characteristics are curves. With have a set of n independent partial differential equations verified by the components of v:

$$\frac{\partial v_i}{\partial t} + \lambda_i \frac{\partial v_i}{\partial x} = 0\,, \qquad i = 1, \ldots, n\,. \tag{3.87}$$

where the λ_i are the diagonal coefficients of Λ. The sets of solutions to each equations is given by:

$$v_i(x,t) = v_i(x - \lambda_i t, 0)\,, \tag{3.88}$$

and the initial condition $v(\cdot,0) = P^{-1}u_0(\cdot)$ will provide each solution $v_i(\cdot,0)$.

Remark 3.15. In a space of dimension $d > 1$, we will consider the system:

$$u_t + \sum_{j=1}^{d} A_j(u)u_j = 0\,. \tag{3.89}$$

and call it as hyperbolic if and only if the matrix $A(u,\xi) = \sum_{j=1}^{d} \xi_j A_j(u)$ is diagonizable and has all real eigenvalues.

3.4.2 Finite difference schemes

To solve numerically the Cauchy problem (3.79), we introduce a uniform discretization of the domain $[0,L] \times \mathbb{R}_+$, of steps h in space and Δt in time, with $h = L/N$, N denoting the number of grid points in $]0,L]$. For $j \in \{1,\ldots,N\}$ (or $j \in \mathbb{Z}$) and $n \in \mathbb{N}$, we denote

$$x_j = jh\,, \qquad \text{and} \qquad t^n = n\Delta t\,.$$

A simple, if not the simplest, finite difference scheme for the advection equation is the *backward decentered* scheme defined as:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{h} = 0\,, \qquad \text{for } j \in \{1,\ldots,N\}, n \geq 0\,. \tag{3.90}$$

We pose $u_j^0 = u_0(x_j)$, $j \in \{1,\ldots,N\}$ to initialize the numerical scheme. The *periodicity* hypothesis is taken into account by imposing the condition:

$$u_0^n = u_N^n \,.$$

Another simple scheme is the *forward decentered* scheme, written as:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_j^n}{h} = 0\,, \qquad \text{for } j \in \{1,\ldots,N\}, n \geq 0\,. \tag{3.91}$$

By taking the mean value of the previous schemes, we obtain the *centered* scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0\,, \qquad \text{for } j \in \{1,\ldots,N\}, n \geq 0\,. \tag{3.92}$$

If we replace in this scheme the value u_j^n by the average value $1/2(u_{j+1}^n - u_{j-1}^n)$, we obtain the well-known *Lax-Friedrichs* scheme:

$$\frac{u_j^{n+1} - \frac{1}{2}(u_{j+1}^n - u_{j-1}^n)}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0\,, \qquad \text{for } j \in \{1,\ldots,N\}, n \geq 0\,. \tag{3.93}$$

and the *Lax-Wendroff* scheme[5] can be written as follows:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2h} - c^2\frac{\Delta t}{2}\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0\,. \tag{3.94}$$

Remark 3.16. All the previous schemes, except for the Lax-Wendroff scheme, were obtained using a first-order Taylor expansion in Δt:

$$u(x_j, t^{n+1}) = u(x_j, t^n) + \Delta t\, u_t(x_j, t^n) + O(\Delta t)\,.$$

Using the advection equation $u_t + cu_c = 0$, this expansion can be written as follows:

$$u(x_j, t^{n+1}) = u(x_j, t^n) - c\,\Delta t\, u_x(x_j, t^n) + O(\Delta t)\,,$$

then, by replacing u_x by the central difference

$$\frac{1}{2h}(u(x_{j+1}, t^n) - u(x_{j-1}, t^n))\,,$$

and by neglecting the term $O(\Delta t)$, we obtain the centered scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0\,. \tag{3.95}$$

The Lax-Wendroff scheme is obtained truncating the Taylor expansion at the second order in Δt:

$$u(x_j, t^{n+1}) = u(x_j, t^n) + \Delta t\, u_t(x_j, t^n) + \frac{\Delta t^2}{2} u_{tt}(x_j, t^n) + O(\Delta t^2)\,.$$

[5] P.D. Lax and B. Wendroff (1960), Systems of conservation laws, *Comm. Pure and Appl. Math.*, **13**, 217-237.

Since $u_t = -c\,u_x$, we have $u_{tt} = -c\,u_{xt} = -c\,u_{tx} = -c\,(-c\,u_x)_x = c^2\,u_{xx}$ and then we have:

$$u(x_j, t^{n+1}) = u(x_j, t^n) - c\,\Delta t\,u_x(x_j, t^n) + c^2\frac{\Delta t^2}{2}u_{xx}(x_j, t^n) + O(\Delta t^2)\,.$$

Intoducing the approximations:

$$\begin{aligned} u_x &\approx \frac{u(x_{j+1}, t^n) - u(x_{j-1}, t^n)}{2h} \\ u_x x &\approx \frac{u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)}{h^2}\,, \end{aligned}$$

and neglecting the $O(\Delta t^2)$ term, we obtain the finite difference approximation:

$$\begin{aligned} \frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\Delta t} &\approx c\,\frac{u(x_{j+1}, t^n) - u(x_{j-1}, t^n)}{2h} \\ &\quad + c^2\,\frac{\Delta t}{2}\frac{u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)}{h^2}\,. \end{aligned}$$

and the Lax-Wendroff scheme is directly derived from this approximation.

We can also consider the *three-levels, leap-frog* scheme (similar to the Richardson scheme for the heat equation):

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0\,. \tag{3.96}$$

The *centered implicit* scheme is defined as:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2h} = 0\,. \tag{3.97}$$

and the *totally centered implicit* scheme (cf. Crank-Nicolson's scheme) as follows:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{2}\left(\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2h} + \frac{u_{j+1}^n - u_{j-1}^n}{2h}\right) = 0\,. \tag{3.98}$$

3.4.3 Consistent scheme

In this section, we focus on the backward decentered scheme (3.90).

Definition 3.10 (consistency). *The backward decentered scheme (3.90) is said to be* consistent *with the partial differential operator $\partial_t + c\partial_x$, if for any sufficiently smooth function $u = u(x,t)$, the difference denoted $(\varepsilon_h u)(x,t)$:*

$$\left(\frac{u(x, t+\Delta t) - u(x,t)}{\Delta t} + c\frac{u(x,t) - u(x-h,t)}{h}\right) - (u_t + c\,u_x)(x,t) \tag{3.99}$$

Numerical scheme	Truncation error
backward decentered	$O(h + \Delta t)$
forward decentered	$O(h + \Delta t)$
centered	$O(h^2 + \Delta t)$
Lax-Friedrichs	$O\left(\dfrac{h^2}{\Delta t} + \Delta t\right)$
Lax-Wendroff	$O(h^2 + \Delta t^2)$
Leap-frog	$O(h^2 + \Delta t^2)$
centered implicit	$O(h^2 + \Delta t)$
totally centered	$O(h^2 + \Delta t^2)$

Table 3.2. *Order of the truncation errors related to various numerical schemes used to solve the 1d advection equation $u_t + c\,u_x = 0$.*

tends toward 0 when h and Δt tend toward 0 independently.
Moreover, if there exists a constant $C > 0$, independent of u and of its derivatives, such that $(\varepsilon_h u)(x,t)$ tends to 0 as $O(h^p + \Delta t^q)$ when $h \to 0$, $\Delta t \to 0$:

$$\|\varepsilon_h u\|_\infty \leq C(h^p + \Delta t^q)\,, \qquad p > 0, q > 0\,, \tag{3.100}$$

the scheme is said to be accurate *at the order p in space and at the order q in time.*

Proposition 3.10. *The backward decentered scheme (3.90) is* consistent *with the advection equation and is* first-order accurate *in space and in time.*

Proof. A Taylor expansion gives easily the result. For any C^2 continuous function v, we have:

$$\frac{v(x,t+\Delta t) - v(x,t)}{\Delta t} + c\frac{v(x,t) - v(x-h,t)}{h} - (v_t + c\,v_x)(x,t)$$
$$= \frac{\Delta t}{2} v_{tt}(x, t+\theta_t) - c\frac{h}{2} v_{xx}(x - \theta_x, t)\,, \quad 0 \leq \theta_t \leq \Delta t\,, 0 \leq \theta_x \leq h\,.$$

The table 3.2 reports the orders of the truncation errors for the schemes we have presented.

3.4.4 Maximum principle

At first, let us notice that according to the expression of the exact solution (3.80), $u(x,t) = u_0(x - ct)$, it is easy to see that this solution satisfies the maximum principle as:

$$\inf_{\xi\in\mathbb{R}} u_0(\xi) \le u(x,t) \le \sup_{\xi\in\mathbb{R}} u_0(\xi)\,, \quad \forall (x,t) \in \mathbb{R}\times\mathbb{R}_+\,. \tag{3.101}$$

We will show that this is also true for the numerical solution u_j^n obtained with the explicit scheme.

Lemma 3.9 (discrete maximum principle). *Let (u_j^n) be the numerical solution obtained using the backward decentered scheme (3.90). Under the assumption:*

$$|\nu| = \left|\frac{c\,\Delta t}{h}\right| \le 1\,, \tag{3.102}$$

we have:

$$\inf_{i\in\mathbb{Z}} u_i^0 \le u_j^n \le \sup_{i\in\mathbb{Z}} u_i^0\,, \quad \forall j\in\mathbb{Z}, \forall n\in\mathbb{N}\,. \tag{3.103}$$

Remark 3.17. The assumption involves the *Courant number* $\nu = c\Delta t/h$. The integer part of $|\nu|$ corresponds to the number of grid points traversed between the times t^n and t^{n+1} by the characteristic curve passing through x_j at time t^{n+1} (Figure 3.1).

If $\nu \in \mathbb{N}$, the point $(x_j - c\,\Delta t, t^n)$ is a grid point; if $\nu = k \in \mathbb{Z}$, we have $x_j - c\,\Delta t = x_{j-k}$ and the exact solution of the advection equation is such that:

$$u(x_j, t^{n+1}) = u(x_{j-k}, t^n)\,, \qquad \forall j\in\mathbb{Z}, \forall n\in\mathbb{N}$$

and hence the numerical scheme $u_j^{n+1} = u_{j-k}^n$ is *exact*.

If $|\nu|$ is not an integer, we introduce $k\in\mathbb{Z}$ such that $k \le \nu < k+1$. The point $x_j - c\Delta t \in [x_{j-k-1}, x_{j-k}]$ and we have:

$$x_j - c\Delta t = (1-(\nu-k))x_{j-k} + (\nu-k)x_{j-k-1}\,.$$

It is then possible to compute an approximation of $u(x_j - c\Delta t, t^n)$ using a linear interpolation in the interval $[x_{j-k-1}, x_{j-k}]$ as follows:

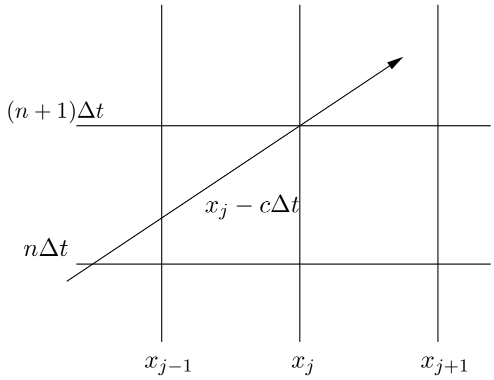

Fig. 3.1. *Characteristic line passing at (x_j, t^{n+1}).*

$$u(x_j - c\Delta t, t^n) \approx (1-(\nu-k))u(x_{j-k}, t^n) + (\nu-k)u(x_{j-k-1}, t^n)\,, \quad (3.104)$$

and we can consider the relevant numerical scheme, for $k \le \nu = c\frac{\Delta t}{h} < k+1$, $k \in \mathbb{Z}$:

$$u_j^{n+1} = (1-(\nu-k))u_{j-k}^n + (\nu-k)u_{j-k-1}^n\,, \quad (3.105)$$

called the *characteristic scheme* with linear interpolation on grid points.

Proposition 3.11. *The advection equation $u_t - c\,u_c = 0$ is not modified by changing simultaneously c in $-c$ and x in $-x$. These changes lead to consider the forward decentered scheme (3.91) instead of the backward decentered scheme (3.90).*

Proof. We assume $c > 0$ and we consider the backward decentered scheme (3.90):

$$u_j^{n+1} = (1-\nu)u_j^n + \nu u_{j-1}^n\,.$$

Under the hypothesis $|\nu| \le 1$, the coefficients $1-\nu$ and ν are either strictly positive or null, and we can deduce easily that:

$$\min(u_j^n, u_{j-1}^n) \;\le\; u_j^{n+1} \;\le\; \max(u_j^n, u_{j-1}^n)\,,$$

and the result is obtained using a recurrence formula. □

Lemma 3.10. *Suppose that the Courant number ν is such that $|\nu| \le 1$. Then the characteristic scheme (3.105) coincide with the backward decentered scheme (3.90) if $c > 0$ and with the forward decentered scheme (3.91) if $c < 0$.*

Remark 3.18. According to this result, we will simply call *upwind scheme* either the backward decentered scheme (if $c > 0$) or the forward decentered scheme (if $c < 0$). The upwind scheme takes always into account the information in the direction from which emanates the characteristics passing through (x_j, t^{n+1}); it can be written as:

$$u_j^{n+1} = \begin{cases} u_j^n - c\dfrac{\Delta t}{h}(u_j^n - u_{j-1}^n)\,, & \text{if } c > 0\,, \\ u_j^n - c\dfrac{\Delta t}{h}(u_{j+1}^n - u_j^n)\,, & \text{if } c < 0\,. \end{cases} \quad (3.106)$$

3.4.5 Convergence

We are now ready to enounce the following convergence result.

Theorem 3.5 (convergence of the upwind scheme (3.106)). *Suppose $u_0 \in C^2(\mathbb{R})$ is periodic of periof $L > 0$ and let u be the exact solution of the advection problem (3.79). Let (u_j^n) denote the solution obtained using the upwind scheme (3.106), with $u_j^0 = u_0(x_j)$ for every $j \in \mathbb{Z}$. Assume also that the Courant number satisfies the condition:*

$$|\nu| = \left| c\frac{\Delta t}{h} \right| \leq 1\,. \tag{3.107}$$

Then, for every $T > 0$, there exists a constant $C_T > 0$ (depending on T and u_0 only) such that:

$$\sup_{\substack{1\leq j\leq N\\ 0\leq t^n\leq T}} |u_j^n - u(x_j, t^n)| \leq C_T\,(h + \Delta t)\,. \tag{3.108}$$

Proof. (similar to the proof of Theorem 3.3). We consider the case $c > 0$ and the backward decentered scheme (3.90). Since the exact solution $u(x,t) = u_0(x - ct)$ is C^2 on $\mathbb{R} \times \mathbb{R}_+$, we obtain thanks to Proposition 3.10:

$$\begin{aligned} &\frac{u(x_j, t^{n+1} - u(x_j, t^n)}{\Delta t} + c\frac{u(x_j,t^n) - u(x_{j-1},t^n)}{h} = \\ &\qquad \frac{\Delta t}{2} u_{tt}(x_j, t^n + \lambda_j^n) - c\frac{h}{2} u_{xx}(x_j - \mu_j^n, t^n)\,, \end{aligned} \tag{3.109}$$

with $0 \leq \lambda_j^n \leq \Delta t$ and $0 \leq \mu_j^n \leq h$. By denoting ε_j^n the right-hand side and $K = \sup_{0\leq x\leq L} |(u_0)''(x)|$ and since

$$u_{xx}(x,t) = (u_0)''(x - ct)\,, \qquad u_{tt}(x,t) = c^2 (u_0)''(x - ct)\,,$$

we obtain the following estimate:

$$|\varepsilon_j^n| \leq |c|\frac{K}{2}h + c^2\frac{K}{2}\Delta t\,. \tag{3.110}$$

We introduce the *numerical error* $e_j^n = u_j^n - u(x_j, t^n)$ and we obtain by difference:

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} + c\frac{e_j^n - e_{j-1}^n}{h} = -\varepsilon_j^n\,. \tag{3.111}$$

Hence,

$$e_j^{n+1} = (1-\nu)e_j^n + \nu e_{j-1}^n - \Delta t \varepsilon_j^n$$

with $\nu = c\Delta t/h$, the Courant number. For $n \geq 0$, we denote

$$E^n = \sup_{1\leq j\leq N} |e_j^n$$

and since $1 - \nu \geq 0$, $\nu \geq 0$, we deduce that:

$$|e_j^{n+1}| \leq E^n + \Delta t\left(|c|\frac{K}{2}h + c^2\frac{K}{2}\Delta t \right)\,,$$

and thus we have:

$$E^{n+1} \leq E^n + \Delta t|c|\frac{K}{2}(1 + |c|)(h + \Delta t)\,, \quad \text{with } E^0 = 0(u_j^0 = u_0(x_j))\,,$$

then

$$E^n \leq n\Delta t|c|\frac{K}{2}(1+|c|)(h+\Delta t)\,, \quad \text{with } C_T = |c|\frac{KT}{2}(1+|c|)\,. \tag{3.112}$$

□

The estimate proves that the upwind scheme converges when the Courant number is (in absolute value) lesser than or equal to one, and we have then:

$$\lim_{h,\Delta t\to 0}\left(\sup_{\substack{1\le j\le N\\ 0\le t^n\le T}} |u_j^n - u(x_j,t^n)|\right) = 0\,. \tag{3.113}$$

3.4.6 Stability

As we have already mentioned, the Lax's theorem can be invoked also for the linear advection equation. Hence, we can expect here to reduce the convergence analysis of consistent schemes to the study of their stability. We have seen in the previous section that the stability property with respect to the norm $\|\cdot\|_2$ requires computing the eigenvalues of matrices and this can be highly tiresome (and not to mention error prone). Here, we establish the stability of the upwind scheme using the Fourier analysis. When hyperbolic equations describe the motion and evolution of waves, the modulus of $\lambda(k)$ represents the *damping* and the argument is the *dispersion* (*i.e.,* the variation of the wave speed with the frequency) in the scheme.

It is possible to find the solution of the equation (3.79) by using the Fourier analysis. Hence, the Fourier mode:

$$u(x,t) = \exp(i(kx+\omega t))\,, \tag{3.114}$$

is an exact solution of the advection equation if $\omega = -ck$, that represents a *dispersion relation.*

Now, we want to compare the continuous solution to the discrete solution obtained using the upwind scheme. To this end, we recall that the approximation generated by the upwind scheme satisfies:

$$u_j^{n+1} = (1-\nu)u_j^n + \nu u_{j-1}^n\,, \qquad \forall j\in\{1,\dots,N\}\,, \tag{3.115}$$

This explicit difference scheme can be equivalently rewritten in the form:

$$u^{n+1} = A_N u^n = (1-\nu)\,I + \nu E^{-1}\,, \tag{3.116}$$

where $u^n = (u_1^n,\dots,u_N^n)^t \in \mathbb{R}^N$, $I\in M_N(\mathbb{R})$ is the identity matrix and the matrices A_N and E^{-1} are respectively defined as:

$$A_N = \begin{pmatrix} 1-\nu & 0 & & & 0 & \nu \\ \nu & 1-\nu & & & & 0 \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \nu & 1-\nu & \\ 0 & & & & \nu & 1-\nu \end{pmatrix}, \quad E^{-1} = \begin{pmatrix} 0 & & & 0 & 1 \\ 1 & 0 & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ 0 & & & 1 & 0 \end{pmatrix},$$

and the eigenvalues λ_k of A_N are given by:

$$\lambda_k = 1 - \nu + \nu \exp\left(-\frac{2ik\pi}{N}\right), \qquad 1 \leq k \leq N\,. \tag{3.117}$$

The coordinates of the eigenvectors $(X_k)_{1\leq k\leq N}$ of A_N corresponding to the eigenvalues λ_k are:

$$(X_k)_j = \frac{1}{\sqrt{N}} \exp\left(\frac{2ijk\pi}{N}\right), \qquad 1 \leq k \leq N\,. \tag{3.118}$$

Proposition 3.12. *Let consider $u = (u_j)_{1\leq j\leq N}$ and $v = (v_k)_{1\leq k\leq N}$ two vectors of $\mathbb{C}^N$ such that:*

$$u_j = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} v_k \exp\left(\frac{2ijk\pi}{N}\right), \qquad 1 \leq j \leq N\,. \tag{3.119}$$

Then v is given by the discrete inverse Fourier transform:

$$v_k = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} u_j \exp\left(-\frac{2ijk\pi}{N}\right), \qquad 1 \leq k \leq N\,, \tag{3.120}$$

and we have:

$$\sum_{k=1}^{N} |v_k|^2 = \sum_{j=1}^{N} |u_j|^2\,. \tag{3.121}$$

Lemma 3.11. *Suppose $0 \leq c\Delta t/h \leq 1$, then the backward decentered numerical scheme (3.90) converges in L^2-norm: if the solution u of the problem (3.79) is sufficiently smooth, then for every $T > 0$:*

$$\lim_{h,\Delta t\to 0} \left(\sup_{t^n\leq T} \|e^n\|_2\right) = 0\,, \tag{3.122}$$

where $e_j^n = u_j^n - u(x_j, t^n)$ and $e^n = (e_j^n)_{1\leq j\leq N}$.

Proof. Considering the eigenvalues λ_k of A_N, it is easy th show that:

$$|\lambda_k|^2 = 1 + 2\nu(\nu - 1)\left(1 - \cos\left(\frac{2k\pi}{N}\right)\right),$$

and thus $|\lambda_k|^2 \leq 1$ for every index k, if $0 \leq \nu \leq 1$. We observe that $v_k^n = (\lambda_k)^n v_k^0$, where the v_k^n are the components of u^n in the basis $(X_k)_{1\leq k\leq N}$ and thus we have, thanks to the identity (3.121):

$$h\sum_{j=1}^{N} |u_j^n|^2 = h\sum_{k=1}^{N} |v_k^n|^2 \leq h\sum_{k=1}^{N} |v_k^0|^2 = h\sum_{j=1}^{N} |u_j^0|^2\,,$$

and this estimate indicates that the numerical scheme (3.90) is stable in L^2-norm and thus this scheme is convergent with respect to this norm according to the Lax's theorem. □

Remark 3.19. We shall notice here that the backward decentered scheme (3.90) converges in L^2-norm if and only if $0 \leq \nu \leq 1$. Indeed, if $\nu \notin [0,1]$ then the relation $|\lambda_k|^2 \leq 1$ is no longer valid for all indices k (only $\lambda_N = 1$) and the numerical scheme cannot be convergent.

Exercise 3.3. Show that the Lax-Wendroff scheme (3.94) converges in L^2-norm if and only if the Courant number is such that $\nu \leq 1$.

Exercise 3.4. Show that the centered scheme (3.95) is *unconditionally unstable* in L^2-norm, *i.e.,* for any value of the Courant number ν.

3.5 The finite difference method for other 1d models

Other boundary value models can be successfully discretized using the finite difference method. For example, we consider the 1d *wave equation* posed in the bounded domain $\Omega =]0,1[$ endowed with the boundary conditions:

$$(\mathcal{W})\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,t) - \dfrac{\partial^2 u}{\partial x^2}(x,t) = 0\,, & \text{for } (x,t) \in \Omega \times \mathbb{R}_+ \\ u(0,t) = u(1,t) = 0\,, & \text{for } t \in \mathbb{R}_+^* \\ u(x,0) = u_0(x)\,, \quad \dfrac{\partial u}{\partial t}(x,0) = u_1(x)\,, & \text{for } x \in \Omega \end{cases} \tag{3.123}$$

where u_0 and u_1 are the given initial data. This model is used to represent the motion of a vibrating string. In this case, the unknown function u may represent the transverse displacement of an elastic vibrating string of length one, attached at the two endpoints. The string could be subjected to a vertical force of density f specified then as a right-hand side term. The function u_0 and u_1 denote respectively the initial displacement and the initial velocity of the string.

3.5.1 General solution

The general solution $u : \mathbb{R}\times]0,T[\to \mathbb{R}$ to the one-dimensional scalar wave equation:

$$\frac{\partial^2 u}{\partial t^2}(x,t) = -c^2 \frac{\partial^2 u}{\partial x^2}(x,t)\,, \qquad x \in \mathbb{R}\,, t \in]0,T[\,, \tag{3.124}$$

with the initial conditions:

$$u(x,0) = u_0(x)\,, \qquad \frac{\partial u}{\partial t}(x,0) = u_1(x)\,, \quad x \in \mathbb{R} \tag{3.125}$$

was already known by the French mathematician and philosopher Jean le Rond d'Alembert (1717-1783). The result is the d'Alembert formula:

$$u(x,t) = \frac{u_0(x-ct)+u_0(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(y)\,dy\,. \tag{3.126}$$

In the classical sense, if $u_0 \in C^k$ and $u_1 \in C^{k-1}$ then $u(x,t) \in C^k$. We observe that the solution $u(x,t)$ depends only on the initial data on the interval $[x-ct, x+ct]$. If the initial data u_0 and u_1 are generalized functions or functions with a compact support included in an interval $[a,b] \subset \mathbb{R}$, then the solution $u(\cdot,t)$ at time t will have its support included in the interval $[a-ct, b+ct]$. The initial data travel at a finite speed c. This is indeed a big conceptual difference with the heat transfer equation described previously, for which the propagation speed was infinite.

Remark 3.20. Suppose that periodic boundary conditions are specified to the model, *i.e.,* $u(x+1,t) = u(x,t)$ for $(x,t) \in \Omega\times\mathbb{R}_+$. If the value of the unknown function u is not fixed at the endpoints of Ω by the boundary conditions, the function u may not remain bounded in time. To overcome this problem, we can assume that the initial velocity is null on average, *i.e.,* that:

$$\int_0^1 u_1(x)\,dx = 0\,. \tag{3.127}$$

3.5.2 Finite difference schemes

We shall proceed like for the heat equation, for which we introduced an approximation of the spatial derivative $\partial^2/\partial x^2$ and use a similar approximation also for the second-order derivative with respect to time. We consider $N+1$ equidistributed grid points in space, $x_j = jh$ and we denote by h the space step, *i.e.,* $h = 1/(N+1)$ and by Δt the time step. The discrete solution u at each time step is a vector $u^n = (u_j^n)_{0\leq j\leq N} \in \mathbb{R}^N$. The boundary conditions impose:

$$u_0^n = u_{N+1}^n = 0\,, \qquad \text{for all } n \geq 0\,.$$

Using the yet classical finite difference approximations of the second-order derivatives $\partial^2/\partial t^2$ and $\partial^2/\partial x^2$ in the equation (3.123), we can easily obtain the *explicit* scheme, for $n \geq 0$ and $0 \leq j \leq N$:

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0\,, \tag{3.128}$$

and the initial conditions are discretized as follows:

$$u_j^0 = u_0(x_j) \quad \text{and} \quad \frac{u_j^1 - u_j^0}{\Delta t} = u_1(x_j)\,, \tag{3.129}$$

Remark 3.21. In the case of periodic boundary conditions, the first-order derivative with respect to time $\partial/\partial t$ would be discretized as follows:

$$\frac{u_j^1 - u_j^0}{\Delta t} = \int_{x_{j-1/2}}^{x_{j+1/2}} u_1(x)\, dx\,,$$

in order for the discrete initial speed to satisfy the condition (3.127) as well.

In order to write the finite difference scheme in a more concise form, we introduce the tridiagonal Laplacian matrix $A_h^{(0)} \in M_N(\mathbb{R})$ defined by (3.15) and then we write in a vector form:

$$u^{n+1} = (2I - (\Delta t)^2 A_h^{(0)})u^n - u^{n-1}\,, \tag{3.130}$$

where I is the identity matrix in $\mathbb{R}^N$.

Similarly, we could have considered the *implicit* scheme defined as:

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} = 0\,, \tag{3.131}$$

with the same discretization of the initial and boundary conditions.

However, the wave equation (3.123) is usually solved using the following centered θ-scheme, $0 \le \theta \le 1/2$:

$$\begin{aligned} &\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \theta\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} \\ -(1-2\theta)&\frac{u_{j+1}^{n} - 2u_j^{n} + u_{j-1}^{n}}{h^2} - \theta\frac{u_{j+1}^{n-1} - 2u_j^{n-1} + u_{j-1}^{n-1}}{h^2} = 0\,, \end{aligned} \tag{3.132}$$

with the initial conditions (3.129). The scheme corresponds to the explicit scheme (3.128) if $\theta = 0$ and is implicit when $\theta \neq 0$.

3.5.3 The von Neumann stability

The stability with respect to the $\|\cdot\|_2$ norm of the explicit scheme (or any of the previous schemes) could be analyzed using the matrix form of the scheme, like we did for the heat equation. However, this analysis would only give us a necessary condition for the stability. Hence, we will study the stability property using the Fourier analysis. We give now a result for the explicit scheme.

Lemma 3.12. *The explicit scheme (3.128) is stable in L^2 norm under the CFL condition:*

$$\frac{\Delta t}{h} \le 1\,. \tag{3.133}$$

Proof. Using the Fourier transformation with respect to the space variable, we can write:

$$\frac{\hat{u}^{n+1}(k) - 2\hat{u}^n(k) + \hat{u}^{n-1}(k)}{\Delta t^2} - \frac{\exp(ihk) - 2 + \exp(-ihk)}{h^2}\hat{u}^n(k) = 0\,,$$

or equivalently:

$$\frac{\hat{u}^{n+1}(k) - 2\hat{u}^{n}(k) + \hat{u}^{n-1}(k)}{\Delta t^2} + \frac{4}{h^2}\sin^2\left(\frac{hk}{2}\right)\hat{u}^{n}(k) = 0\,. \tag{3.134}$$

We introduce the vector:

$$v^n(k) = \begin{pmatrix} u^{n+1}(k) \\ u^n(k) \end{pmatrix}$$

and the three-levels scheme can be expressed also as:

$$\hat{v}^n(k) = \begin{pmatrix} \hat{u}^{n+1}(k) \\ \hat{u}^n(k) \end{pmatrix} = \begin{pmatrix} 2 - a^2(k) & -1 \\ 1 & 0 \end{pmatrix} \hat{v}^{n-1}(k) = B(k)\hat{v}^{n-1}(k)\,, \tag{3.135}$$

with the coefficient $a(k)$ corresponding to:

$$a(k) = 2\frac{\Delta t}{h}\sin\left(\frac{\pi hk}{2}\right)\,.$$

The eigenvalues λ_1, λ_2 of the matrix $B(k)$ are the roots of the second-degree polynomial:

$$\lambda^2 - \lambda(2 - a^2(k)) + 1 = 0\,,$$

that has the following discriminant: $\Delta(k) = a^2(k)\,(a^2(k) - 4)$. In the case where $a(k) = 0$, the matrix $B(k)$ is not normal, neither can be diagonalized. A *necessary condition* to establish the Von Neumann stability is that:

$$\rho(B(k)) < 1\,.$$

Actually, the particular solutions of the wave equation,

$$u_k(x,t) = \sin(k\pi x)\exp(\pm ik\pi t) \tag{3.136}$$

will always have the property that

$$|u_k(x,t)| \le 1\,.$$

It seems then reasonable to ask the numerical solutions to have the same property:

$$|\lambda| \le 1\,.$$

The condition $\rho(B(k)) \le 1$ is equivalent to the following: for every $k \in \mathbb{R}$, $\Delta(k) \le 0$. Now, $\Delta(k) \le 0$ if and only if $|a(k)| \le 2$ and this condition is satisfied, for every $k \in \mathbb{R}$ if and only if:

$$\frac{\Delta t}{h} \le 1\,.$$

□

We have a similar CFL stability condition for the θ-scheme.

Lemma 3.13. *(i) If $1/4 \leq \theta \leq 1/2$, then the θ-scheme (3.132) is* unconditionally stable *in L^2-norm.*

(ii)If $0 \leq \theta < 1/4$, the θ-scheme is stable in L^2-norm under the CFL condition:

$$\frac{\Delta t}{h} < \sqrt{\frac{1}{1-4\theta}}\,. \tag{3.137}$$

and is unstable otherwise.

Proof. (result admitted here).

3.6 The finite difference method in two dimensions

To conclude our presentation of the mathematical properties of finite difference methods for solving boundary-value probems, we shall see how they extend to more space dimensions. We focus here on linear second-order elliptic differential operators in two dimensions of space. In this respect, the Poisson's equation is a fundamental partial differential equation which can be encountered in many applications, for example electrostatics, mechanical engineering and theoretical physics.

We recall that the Poisson's boundary-value problem, endowed with Dirichlet boundary conditions, takes the form:

$$\begin{cases} -\Delta u(x,y) = f(x,y)\,, & (x,y) \text{ in } \Omega \\ u(x,y) = g(x,y)\,, & (x,y) \text{ on } \partial\Omega \end{cases} \tag{3.138}$$

where $\Omega \subset \mathbb{R}^2$ is assumed to be a bounded, connected and open domain and $\partial\Omega$ denotes its boundary. If we consider the parabolic equation:

$$\frac{\partial u}{\partial t}((x,y),t) = \Delta u((x,y),t) + f(x,y)$$

it is easy to see that, if the solution u converges to a limit when t tends to infinity, this limit will be a solution of equation (3.138) corresponding to a steady-state.

3.6.1 Existence and property of a regular solution

We consider here the domain Ω to be the unit square:

$$\Omega = \{(x,y)\,, 0 < x, y < 1\}\,,$$

but the analysis below can be adapted to any rectangle. The following result states that a function u such that $-\Delta u \geq 0$ (resp. $-\Delta u \leq 0$) attains its absolute minimum (resp. maximum) on the boundary $\partial\Omega$.

Lemma 3.14 (maximum principle). *Consider $u \in C^2(\Omega)$ such that*

$$-\Delta u \geq 0\,, \ \textit{in } \Omega \qquad \textit{and} \qquad u \geq 0\,, \ \textit{on } \partial\Omega\,.$$

Then $u \geq 0$ in $\bar{\Omega}$.

Proof. (i) Let consider $v \in C^2(\Omega)$ a function which attains its minimum value at an internal point $(x_0, y_0) \in \Omega$. The linear mapping $v_{y_0} :]0,1[\to \mathbb{R}$, $t \mapsto (t, y_0)$ has a relative minimum at $t = x_0$. Using the Taylor-Lagrange formula, we can easily deduce that:

$$\frac{\partial^2 v_{y_0}}{\partial t^2}(x_0) \geq 0\,,$$

which is equivalent to

$$\frac{\partial^2 v}{\partial x^2}(x_0, y_0) \geq 0\,.$$

Similarly, we have:

$$\frac{\partial^2 v}{\partial y^2}(x_0, y_0) \geq 0\,,$$

and adding these two inequalities yields:

$$\Delta v(x_0, y_0) \ \geq \ 0\,.$$

(ii) Consider $u \in C^2(\Omega)$ such that $-\Delta u \geq 0$ in Ω and $u \geq 0$ on $\partial\Omega$. Suppose that $u(x) < 0$ in $\bar{\Omega}$. Since $\bar{\Omega}$ is a compact subset and u is a continuous function, u attains its absolute (negative) minimum at $(x_0, y_0) \in \bar{\Omega}$. However, $(x_0, y_0) \notin \partial\Omega$ as $u \geq 0$ on $\partial\Omega$ by assumption. Let us denote the minimum by:

$$-M = u(x_0, y_0) = \min_{\bar{\Omega}} u(x, y) < 0\,.$$

Now, we introduce the auxiliary function, continuous on the compact $\bar{\Omega}$:

$$u_\varepsilon(x, y) = u(x) - \varepsilon(x^2 + y^2)\,, \quad \text{with} \quad 0 < \varepsilon < \ \frac{M}{2}\,,$$

it attains its absolute minimum at the point (x_1, y_1). We have then at the point (x_0, y_0):

$$u_\varepsilon(x_0, y_0) = -M - \varepsilon(x_0^2 + y_0^2) \ < \ -M$$

and then $u_\varepsilon(x_1, y_1) \ \leq \ u_\varepsilon(x_0, y_0) \ < \ -M$ and $u(x, y) \geq 0$ if $(x, y) \in \partial\Omega$. Hence, we have:

$$u_\varepsilon(x, y) \ \geq \ -\varepsilon(x^2 + y^2) \ \geq \ -2\varepsilon \ > \ -M\,.$$

Now, since $u_\varepsilon(x_1, y_1) \ < \ u_\varepsilon(x, y)$ for all $(x, y) \in \partial\Omega$, we deduce that $(x_1, y_1) \notin \partial\Omega$ and thus $(x_1, y_1) \in \Omega$. If we apply the result obtained in *(i)* to u_ε at the point (x_1, y_1), we find that:

$$-\Delta u_\varepsilon(x_1, y_1) \leq 0\,.$$

However, we know that $-\Delta u \geq 0$ in Ω and thus

$$-\Delta u_\varepsilon(x, y) = -\Delta u(x, y) + \varepsilon \Delta(x^2 + y^2) = -\Delta u(x, y) + 4\varepsilon > 0\,,$$

and in particular at point $(x_1, y_1) \in \Omega$ we have:

$$-\Delta u_\varepsilon(x_1, y_1) > 0$$

this is in clear contradiction with the assumption $u(x) < 0$ in $\bar{\Omega}$ and the results follows easily. □

Corollary 3.1. *Consider $u \in C^2(\Omega)$ such that*

$$-\Delta u = 0\,, \text{ in } \Omega \qquad \text{and} \qquad u = 0\,, \text{ on } \partial\Omega\,.$$

Then $u \equiv 0$ in $\bar{\Omega}$.

Proof. Simply invoke the previous lemma with u and $-u$. □

In order to establish the existence and uniqueness of a solution for the Poisson problem, we shall consider boundary conditions of the form:

$$\begin{aligned} u(0, y) = u(1, y) &= 0\,, && 0 \leq y \leq 1 \\ u(x, 0) &= 0\,, && 0 \leq x \leq 1 \\ u(x, 1) &= g(x)\,, && 0 < x < 1\,,\ g \in C^1(\bar{\Omega}) \end{aligned} \tag{3.139}$$

and we simplify the analysis by considering the homogeneous Laplace problem, *i.e.,* for $f \equiv 0$. Using a method of separation of variables, we can write the solution as $u(x, y) = X(x)Y(y)$ and considering the Laplace problem leads to solve the following equation:

$$X''(x)Y(y) + X(x)Y''(y) = 0\,, \quad \text{or} \quad -\frac{X''(x)}{X(x)} = \frac{Y''(y)}{Y(y)}\,. \tag{3.140}$$

Both sides have to be equal to a constant, say Λ, independent of x and y variables. We obtain the eigenvalue problem, for the function X:

$$-X''(x) = \lambda X(x)\,, \quad 0 < x < 1\,, \qquad \text{with } X(0) = X(1) = 0\,, \tag{3.141}$$

Here, we conclude that the eigenvalues and eigenvectors are respectively:

$$\lambda_k = (k\pi)^2\,, \qquad \text{and} \qquad X_k(x) = \sin(k\pi x)\,, \qquad k \in \mathbb{N}^*\,.$$

The theory of harmonic functions allows us to write the formal solutions of Laplace's problem as follows:

$$u(x, y) = \sum_{k \geq 1} c_k \sin(k\pi x)\sin(k\pi y)\,, \tag{3.142}$$

where c_k are arbitrary scalar values. If we assume that the function $g(x)$ admits a Fourier sine series:

$$g(x) = \sum_{k\geq 1} g_k \sin(k\pi x)\,, \qquad g_k = 2\int_0^1 g(x)\sin(k\pi x)\,dx\,,$$

then we obtain the value of the coefficients:

$$g_k = g_k/\sin(k\pi)\,, \qquad \text{for } k \geq 1\,.$$

and this provides the complete solution of the Laplace problem with the specific Dirichlet boundary conditions (3.139).

Lemma 3.15 (existence). *The Laplace problem $-\Delta u = 0$ with the boundary conditions (3.139) admits a solution u.*

In addition, we can provide a result on the uniqueness of a solution for the Poisson problem; its existence remains more tedious to establish.

Lemma 3.16 (uniqueness). *The Poisson problem (3.138) has a* unique *solution, if it exists.*

Proof. Assume u_1 and u_2 two solutions of the Poisson problem and pose $v = u_1 - u_2$. Then we have $\Delta v = \Delta u_1 - \Delta u_2 = f - f = 0$ in Ω and $v = g - g = 0$ on $\partial\Omega$. According to Corollary (3.1), we have $v = 0$. □

3.6.2 Finite difference approximations

On a general domain Ω it is not possible to find an analytical expression of the solution of Poisson's equation. Therefore, we resolve to replace this problem by a discrete problem. The unit square Ω is discretized using a set of grid points:

$$\bar{\Omega}_h = \{(x_j, y_k)\,,\quad 0 \leq j,k \leq N+1\}\,,$$

where N defines the spacing $h = 1/(N+1)$ and $x_j = jh$, $y_k = kh$. The set $\bar{\Omega}_h$ contains $(N+2)^2$ grid points, N^2 interior points and $4(N+1)$ points located on the boundary $\partial\Omega_h$.

Next, we shall define a finite difference operator to approximate the differential Laplacian operator.

Definition 3.11. *Let $u \in C^0(\Omega)$. We call five-points* discrete Laplace operator *(or* discrete Laplacian*) of u the value $\Delta_h u$ defined at every interior grid point (x_j, y_k) as:*

$$\Delta_h u(x_j, y_k) = \frac{1}{h^2}\left(-4u_{j,k} + u_{j-1,k} + u_{j+1,k} + u_{j,k-1} + u_{j,k+1}\right)\,, \tag{3.143}$$

where, as usual, $u_{j,k} = u(x_j, y_k)$.

The stencil of the discrete Laplacian involves the four closest neighbours of the grid point (x_j, y_k).

We have the following consistency result.

Lemma 3.17 (consistency). *Suppose $u \in C^4(\bar{\Omega})$. Then, for every indices $1 \leq j, k \leq N$, we have:*

$$|(\Delta u)(x_j, y_k) - (\Delta u_h)(x_j, y_k)| \leq \frac{h^2}{12} \max_{\bar{\Omega}} \left(\left| \frac{\partial^4 u}{\partial x^4} \right| + \left| \frac{\partial^4 u}{\partial y^4} \right| \right) . \tag{3.144}$$

and thus the discrete Laplacian is a consistent second-order approximation of the Laplace operator.

Proof. We introduce the function $\theta_1^{jk}(t) = u(jh + t, kh)$ which is defined and continuous in a neighborhood of 0 in t, *i.e.*, containing at least $]-h, h[$. Hence we have:

$$\frac{\partial^n u}{\partial x^n}(jh + t, kh) = \frac{d^n \theta_1^{jk}}{dt^n}(t) , \qquad 0 \leq n \leq 4 ,$$

Similarly, if we consider the function $\theta_2^{jk}(s) = u(jh, kh + s)$ we can write:

$$\frac{\partial^n u}{\partial y^n}(jh, kh + s) = \frac{d^n \theta_2^{jk}}{ds^n}(s) , \qquad 0 \leq n \leq 4 ,$$

and, in particular we have:

$$\Delta u(x_j, y_k) = (\theta_1^{jk})''(0) + (\theta_2^{jk})''(0) .$$

Using Taylor-Lagrange developments, we obtain:

$$h^2 (\theta_1^{jk})''(0) = \theta_1^{jk}(-h) - 2\theta_1^{jk}(0) + \theta_1^{jk}(h) + \frac{h^4}{12}(\theta_1^{jk})^{(4)}(t_{jk}) , \quad t_{jk} \in]-h, h[$$

$$h^2 (\theta_2^{jk})''(0) = \theta_2^{jk}(-h) - 2\theta_2^{jk}(0) + \theta_2^{jk}(h) + \frac{h^4}{12}(\theta_2^{jk})^{(4)}(s_{jk}) , \quad s_{jk} \in]-h, h[$$

and then it is easy to see that:

$$\begin{aligned} &\theta_1^{jk}(0) = u(x_j, y_k) , \quad \theta_1^{ij}(-h) = u(x_{j-1}, y_k) , \quad \theta_1^{ij}(h) = u(x_{j+1}, y_k) \\ &\theta_2^{jk}(0) = u(x_j, y_k) , \quad \theta_2^{ij}(-h) = u(x_j, y_{k-1}) , \quad \theta_2^{ij}(h) = u(x_j, y_{k+1}) \end{aligned}$$

Thus, we have finally:

$$\Delta u(x_j, y_k) = (\Delta_h u)(x_j, y_k) + \frac{h^2}{12} \left(\frac{\partial^4 u}{\partial x^4}(jh + t_{jk}, kh) + \frac{\partial^4 u}{\partial y^4}(jh, kh + s_{jk}) \right)$$

and the result follows by taking the absolute values. □

The finite difference solution $u_h : \bar{\Omega}_h \to \mathbb{R}$ of the problem (3.139) is such that:

$$\begin{cases} -(\Delta_h u_h)(x_j, y_k) = f(x_j, y_k)\,, & \text{for all } (x_j, y_k) \in \Omega_h \\ u_h(x_j, y_k) = g(x_j, y_k)\,, & \text{for all } (x_j, y_k) \in \partial\Omega_h \end{cases} \tag{3.145}$$

The values $(\Delta_h u_h)(x_j, y_k)$ are then defined for all interior points and the Dirichlet boundary conditions supply the values at the boundary points. This leads to a system of $N \times N$ linear equations in the N^2 unknowns $(u_h(x_j, y_k))_{1 \leq j,k \leq N}$.

Contrary to the one dimensional case, where unknowns and equations are stored "naturally" as the components of a vector, we have here to deal directly with the components of a matrix. Rearranging the values as a column vector raises the delicate issue of grid point renumbering. We agree to note the grid points from "the left to the right" and from "the bottom to the top", *i.e.,* according to the increasing order of j and k indices. Hence, the number corresponding to the point (x_j, y_k) will be: $(k-1)N + j$. Notice that only interior points are considered here. The vector (u_h) is then defined by its components:

$$(u_h)_n = u_h(x_j, y_k)\,, \qquad n = (k-1)N + j\,,\ 1 \leq j, k \leq N\,. \tag{3.146}$$

The discrete problem can be written in the vector form as follows:

$$A_h u_h = b_h\,, \tag{3.147}$$

where $A_h \in M_{N^2}(\mathbb{R})$ is the block tridiagonal matrix defined as:

$$A_h = \frac{1}{h^2}\begin{pmatrix} L_4 & -I & 0 & \dots & & & 0 \\ -I & L_4 & -I & 0 & & & \\ 0 & \ddots & \ddots & \ddots & & & \\ \vdots & & \ddots & & & & 0 \\ & & & & -I & -L_4 & -I \\ 0 & & & & 0 & -I & L_4 \end{pmatrix} \quad \text{with } L_4 = \begin{pmatrix} 4 & -1 & 0 & \dots & 0 \\ -1 & 4 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 4 & -1 \\ 0 & \dots & 0 & -1 & 4 \end{pmatrix} \tag{3.148}$$

the zeros in A_h correspond to the $N \times N$ null matrix and I denotes the $N \times N$ identity matrix. And b_h is the vector of $\mathbb{R}^{N^2}$ defined as:

$$b_h = \begin{pmatrix} f_1 + \frac{1}{h^2}(g(1,0)+g(0,1)) \\ f_2 + \frac{1}{h^2}g(2,0) \\ \vdots \\ f_N + \frac{1}{h^2}(g(N,0)+g(0,N)) \\ f_{N+1} + \frac{1}{h^2}g(0,2) \\ f_{N+2} \\ \vdots \\ f_{2N} + \frac{1}{h^2}g(N+1,2) \\ f_{2N+1} + \frac{1}{h^2}g(0,3) \\ f_{2N+2} \\ \vdots \end{pmatrix} \quad \text{where } f_n = f(x_j,y_k)\,, n=(k-1)N+j\,. \tag{3.149}$$

In view of analyzing the existence and uniqueness of a solution to the discrete problem, we introduce the set D_h of all grid functions defined on $\bar{\Omega}_h$, *i.e.,*:

$$D_h = \{v\,,\ \ v : \bar{\Omega}_h \to \mathbb{R}\}\,, \tag{3.150}$$

and its subset $D_{h,0}$ defined as:

$$D_{h,0} = \{v \in D_h\,,\ \ v|_{\partial\Omega_h} = 0\}\,,$$

that can be useful when considering homogeneous boundary conditions. Furthermore, we define the discrete inner product $\langle\cdot,\cdot\rangle_h$ by:

$$\langle u,v\rangle_h = h^2 \sum_{j,k=0}^{N} u_{j,k}\, v_{j,k}\,, \quad \text{for } u,v \in D_{h,0}\,. \tag{3.151}$$

Then, considering homogeneous Dirichlet conditions, we have:

Lemma 3.18. *The operator Δ_h is symmetric and positive definite,* i.e., *we have:*

$$\langle \Delta_h u,v\rangle_h = \langle u,\Delta_h v\rangle_h\,,\ \forall u,v \in D_{h,0}\,, \quad \textit{and} \quad \langle \Delta_h u,v\rangle_h \ \geq\ 0\,,\ \forall v \in D_{h,0}\,. \tag{3.152}$$

with equality only if $v \equiv 0$.

Proof. It is straightforward to show that:

$$\begin{aligned} \langle \Delta_h u,v\rangle_h = \sum_{j=0}^{N}\sum_{k=0}^{N}&((u_{j+1,k}-u_{j,k})(v_{j+1,k}-v_{j,k}) \\ &+ (u_{j,k+1}-u_{j,k})(v_{j,k+1}-v_{j,k})) \\ = \langle u,&\Delta_h v\rangle_h\,, \end{aligned}$$

for $u,v \in D_{h,0}$, thus proving the symmetry property of the operator Δ_h. It is also easy to see, from the previous identity that:

$$\langle \Delta_h v, v\rangle_h = \sum_{j=0}^{N}\sum_{k=0}^{N}((v_{j+1,k}-v_{j,k})^2+(v_{j,k+1}-v_{j,k})^2) \geq 0\,. \qquad (3.153)$$

From the previous identity, we deduce that if $\langle \Delta_h v, v\rangle_h = 0$ then $v_{j+1,k} = v_{j,k}$ and $v_{j,k+1} = v_{j,k}$, for $0 \leq j,k \leq N$. Since we assumed homogenous boundary conditions, $v_{0,k} = 0$ and $v_{j,0} = 0$ and thus we conclude that $v \equiv 0$. □

This property of the discrete operator Δ_h implies that the discrete system has at most one solution.

Lemma 3.19. *The matrix A_h is symmetric and positive definite and monotone. Hence, the discrete problem (3.145) is a well-posed problem.*

Proof. The symmetry property and the positive definitness property of the matrix are established like in the one dimensional case. We rely on the discrete maximum principle to show the monotonicity of A_h. Consider $v_h \in \mathbb{R}^{N^2}$. We identify v_h with a function $v_h : \Omega_h \to \mathbb{R}$ by posing:

$$v_h(x_j, y_k) = (v_h)_n\,, \qquad n = (k-1)N + j\,.$$

If we extend v_h by 0 on $\bar{\Omega}_h$, we have:

$$(A_h v_h)_n = -(\Delta_h v_h)(x_j, y_k)\,, \quad \text{for } n = (k-1)N + j\,.$$

Suppose $(A_h v_h) \geq 0$, then this implies that $-(\Delta_h v_h)(x_j, y_k) \geq 0$. If the minimum of v_h is attained on $\partial\Omega_h = \bar{\Omega}_h \backslash \Omega_h$, then $v_h \geq 0$. On the other hand, v_h attains its minimum in Ω_h at a point (x_j, y_k) such that:

$$v_h(x_j, y_k) \leq v_h(x_l, y_m)\,, \qquad \forall\, l, m \in \{1, \ldots, N\}\,.$$

At this point, we have: $-\Delta_h v_h(x_j, y_k) \leq 0$ and since $A_h v_h \geq 0$ we deduce that:

$$-\Delta_h v_h(x_j, y_k) = 0\,,$$

and thus $v_h(x_j, y_k) = \min_{\bar{\Omega}_h} v_h$ and we have the identities:

$$v_h(x_j, y_k) = v_h(x_{j-1}, y_k) = v_h(x_{j+1}, y_k) = v_h(x_j, y_{k-1}) = v_h(x_j, y_{k+1})\,.$$

And by repeating gradually, we reach the boundary of Ω_h where $v_h = 0$. Hence $\min_{\bar{\Omega}_h} v_h = 0$ and we deduce also that $v_h \geq 0$. □

Lemma 3.20. *We have: $\|A_h^{-1}\|_\infty \leq 1/2$, or in a vector form:*

$$\|A_h^{-1} b_h\|_\infty \leq \frac{1}{2}\|b_h\|_\infty\,. \qquad (3.154)$$

Proof. By identifying v_h with a function $v_h : \Omega_h \to \mathbb{R}$ that vanishes on the boundary (like previously), we have: $-\Delta_h v_h = b_h$ in Ω_h. We introduce the discrete function:

$$z_h(x_j, y_k) = \frac{h^2}{4} 9j^2 + k^2)\,,$$

such that $-\Delta_h z_h = -1$ in Ω_h, and we pose:

$$\omega_h^+ = \| - \Delta_h v_h \|_\infty z_h - v_h\,.$$

Thus, we have:

$$\begin{aligned} -\Delta_h \omega_h^+ &= -\|\Delta_h v_h\|_\infty \Delta_h z_h + \Delta_h v_h \\ &= \Delta_h v_h - \| - \Delta_h v_h\|_\infty \\ &\leq 0 \end{aligned}$$

Hence, ω_h^+ attains its maximum in $\bar{\Omega}_h \backslash \Omega_h$, however we have $v_h = 0$ on $\bar{\Omega}_h \backslash \Omega_h$ and thus:

$$\omega_h^+(x_j, y_k) \;\leq\; \| - \Delta_h v_h\|_\infty \max_{\bar{\Omega}_h \backslash \Omega_h} z_h = \frac{1}{2} \| - \Delta_h v_h\|_\infty\,,$$

and we deduce that:

$$v_h = \| - \Delta_h v_h\|_\infty z_h - \omega_h^+ \;\geq\; -\omega_h^+ \;\geq\; -\frac{1}{2} \| - \Delta_h v_h\|_\infty\,.$$

We proceed similarly with a function $\omega_h^- = \| - \Delta_h v_h\|_\infty z_h + v_h$ and we deduce that:

$$v_h \;\leq\; \frac{1}{2} \| - \Delta_h v_h\|_\infty\,.$$

These two inequalities lead us to conclude that:

$$\|v_h\|_\infty = \max_{\Omega_h} |v_h| \;\leq\; \frac{1}{2} \| - \Delta_h v_h\|_\infty\,.$$

The vector form is easy to deduce from the previous estimate. □

The following error estimate for the finite difference method is a generalization of Theorem (3.1).

Theorem 3.6 (convergence). *Lt $u \in C^4(\bar{\Omega})$ be the solution of the continuous problem and u_h be the solution of the corresponding discrete problem. Then:*

$$\max_{1 \leq j,k \leq N} |u(x_j, y_k) - u_h(x_j, y_k)| \;\leq\; \frac{h^2}{24} \max_{\bar{\Omega}} \left(\left| \frac{\partial^4 u}{\partial x^4} \right| + \left| \frac{\partial^4 u}{\partial y^4} \right| \right)\,. \qquad (3.155)$$

Proof. (admitted here). □

Remark 3.22. (i) In the previous estimate, a scaling factor α appears:

$$\alpha = \frac{ah^2}{24} = \frac{a^4}{24(N+1)^2}\,,$$

for a square domain $\Omega =]0,a[\times]0,a[$ (in the estimate $a=1$). This factor increases rapidly with a if the number of grid points remains fixed. This emphasizes the rapid degradation of the accuracy if the size a of the domain is large and the number of points is fixed.

(ii) The finite difference method can be extended to three dimensions without any further difficulty. However, the exponential increase of the matrix size jeopardizes this method in more than two dimensions.

3.6.3 Boundary conditions on curved boundaries

The discrete form of the Poisson equation has to be modified in the neighborhood of a boundary which does not lie along the grid lines. For example, we could consider a boundary-value problem posed in a domain Ω with a circular hole (Figure 3.2). The square is covered by a regular grid of size $h = 1/N+1$. Obviously, when the points are near the boundary, we need to apply some specific formula for our finite difference approximations. Two strategies can be employed to overcome this problem. On the one hand, we can enforce the boundary condition at the grid points in $\bar{\Omega}$ close to the boundary. These points are then arbitrarily converted into boundary points. The main advantage of this technique is that it allows to use the five-points stencil of the discrete Laplacian for all other grid points. However, this requires defining a prolongation of the function g inside the domain Ω. Moroever, we introduce an error of order h in dealing with the boundary condition.

On the other hand, we could consider enforcing the boundary condition at the intersection points between the grid and the boundary of the domain Ω. Consider the point P which has three neighbors, on the North, West and East, that are ordinary grid points, but neighbor South is inside the circle (Figure 3.2, right-hand side). Hence, computing the derivative u_{xx} will not pose any problem, however computing u_{yy} would require using the point N and a boundary point B instead of the grid point S.

A straightforward technique to find the desired approximation consists in finding three coefficients α, β and γ such that:

$$\alpha u(x_N,y_N) + \beta u(x_P,y_P) + \gamma u(x_S,y_S) = \frac{\partial^2 u}{\partial y^2}(x_P,y_P) + O(h)\,. \tag{3.156}$$

Using Taylor expansions of $u(x_N,y_S)$ and $u(x_S,y_S)$ in the vicinity of P, we obtain the following system to solve:

$$\begin{cases} \alpha+\beta+\gamma = 0 \\ \alpha h - \gamma h_1 = 0 \\ \alpha\dfrac{h^2}{2} + \gamma\dfrac{h_1^2}{2} = 1\,, \end{cases} \tag{3.157}$$

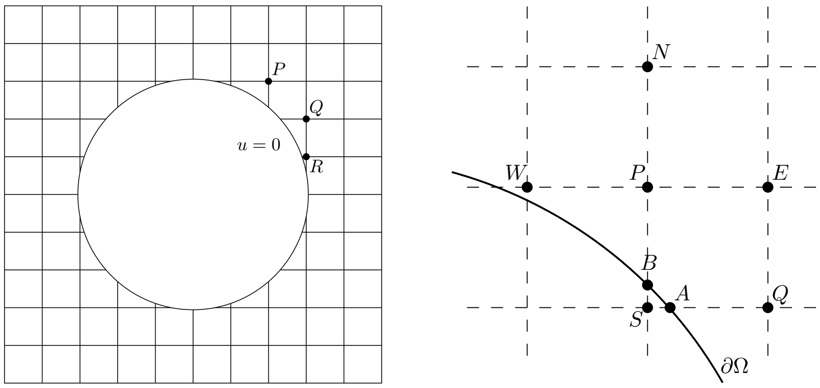

Fig. 3.2. *Boundary-value problem with curved boundary (left) and Dirichlet boundary condition on a curved boundary when the point S is outside the domain Ω (right).*

where h_1 denotes the distance PB. The unique solution to this system is given by:

$$\alpha = \frac{2}{h(h+h_1)}\,, \qquad \beta = -\frac{2}{hh_1}\,, \qquad \gamma = \frac{2}{h_1(h+h_1)}\,. \tag{3.158}$$

Remark 3.23. (i) With this discretization, the truncation error is in $O(h)$ only. Furthermore, we observe that the matrix A_h is no longer symmetric.

(ii) The methods for approximating curved boundaries lead in general to truncation errors of lower order than those at ordinary interior points. Moroever, at the boundary points, the truncation error may not tend to zero when the size tends to zero.

(iii)Normal derivative boundary conditions (Neumann) can be handled in a similar manner, although the technique is much more tedious to carry on. It rquires computing the intersection of the oriented normal vector with the grid lines.

3.6.4 Approximations of other models

The natural generalization in two dimensions of the one-dimensional parabolic heat equation (3.41) is the following equation posed in the unit square domain $\Omega =]0,1[\times]0,1[\subset \mathbb{R}^2$:

$$\begin{cases} u_t - a\,(u_{xx} + u_{yy}) = 0\,, & (x,y) \in \Omega \;\; t > 0\,, \\ \qquad u(x,y,0) = u_0(x,y)\,, & (x,y) \in \Omega\,, \end{cases} \tag{3.159}$$

where a is a positive constant, endowed with Dirichlet boundary conditions, *i.e.*, $u(x,t)$ is given at all points of the boundary of Ω.

We want to approximate the solution $u(x,y,t)$ at discrete grid points (x_j, y_k, t^n). Thus, we will define all spatial grid points: $(x_j, y_k) = (jh, kh)$, where $h = 1/(N+1)$ and the temporal grid points: $t^n = n\Delta t$, for suitably chosen Δt. We already know from theory that we shall satisfy the condition:

$$\Delta t \leq \frac{h^2}{2a} \tag{3.160}$$

in order for the finite difference method to be stable. We proceed as usual for replacing any derivative by finite differences. For any point (x_j, y_k, t^n), we set:

$$u_t \approx \frac{u_{j,k}^{n+1} - u_{j,k}^n}{\Delta t}, \tag{3.161}$$

and we use the discrete Laplacian Δ_h described above to provide the approximations of $u_{xx} + u_{yy}$. The finite difference approximation of the heat equation becomes then:

$$\frac{u_{j,k}^{n+1} - u_{j,k}^n}{\Delta t} - \frac{a}{h^2}\left(-4u_{j,k}^n + u_{j-1,k}^n + u_{j+1,k}^n + u_{j,k-1}^n + u_{j,k+1}^n\right). \tag{3.162}$$

Expressing $u_{j,k}^{n+1}$ leads to:

$$u_{j,k}^{n+1} = u_{j,k}^n + a\frac{\Delta t}{h^2}\left(-4u_{j,k}^n + u_{j-1,k}^n + u_{j+1,k}^n + u_{j,k-1}^n + u_{j,k+1}^n\right). \tag{3.163}$$

The analysis of this scheme is mostly a natural extension of the one-dimensional analysis.

It would be interesting to suggest a two-dimensional extension of the θ-scheme proposed in one dimension. Recall that the advantage of such method was to remove the stability restriction on Δt at a very little computational effort. In two dimensions however, this is not true. The method is still stable without restriction on Δt, but the amount of computational effort is very important. Indeed, it requires solving a system of $(N-1)^2$ linear equations for the unknown values $u_{j,k}^{n+1}$, *i.e.,* at each time step. Each equation involves at most five unknowns, because of the particular stencil we use, but the global matrix is no longer tridiagonal. Moreover, the matrix cannot be transformed in a block diagonal or narrow-band matrix by permuting the rows and columns.

It is surely interesting to search for implicit schemes since they are efficient in one dimension. One idea consists in designing a method which is implicit in one dimensions, but not both. Methods combining two such schemes have been proposed and successfully used[6] in applications.

3.7 Numerical experiments

We present here some numerical results obtained using some of the numerical schemes described in the previous sections.

[6] The first such method was probably the following:
D.W. Peaceman and H.H. Rachford (1955), The numerical solution of parabolic and elliptic di?erential equations, *J. Soc. Indust. Appl. Math.*, **3**, 28.

3.7.1 Resolution of the 1d diffusion problem

Let us consider the following boundary value problem which consists in finding a function u solution of the partial differential equation:

$$-u''(x) = f(x)\,, \qquad 0 < x < 1\,, \tag{3.164}$$

with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$, where the function f is defined as:

$$f(x) = (\pi^2 - 16)u(x) + 8\pi \exp(-4x)\cos(\pi x)\,.$$

The exact solution u to this problem can be easily expressed as:

$$u(x) = \exp(-4x)\sin(\pi x)\,. \tag{3.165}$$

We solved the problem for different grid sizes, *i.e.,* for different values of N. In Figure 3.3 (top), we plotted the exact (solid line) and approximate solutions (dahsed line) for $N = 8$ and $N = 16$.

Since the exact solution is known, it is interesting to estimate the rate of convergence of the finite difference approximation. We computed the error $e_h = (e_j)_{1\leq j\leq N}$ for various grid sizes, in the $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ norms defined in Section 3.2. We estimated the slope of the lines and we found experimentally that the orders of convergence are: $p \approx 1$ for the $\|\cdot\|_1$ norm, $p \approx 1.5$ for the $\|\cdot\|_2$ norm and $p \approx 2$ for the $\|\cdot\|_\infty$ norm, respectively. On the other hand, the theoretical analysis tells us that, if $e_j = u(x_j) - u_j$:

$$\max_{1\leq j\leq N} |e_j| \leq \frac{h^2}{96}\, \|u^{(4)}(x)\|_\infty\,,$$

Hence, it is easy to deduce that $p = 2$ for the $\|\cdot\|_\infty$ norm.

The problem (3.164) can be slightly generalized under the following form:

$$-(a\,u')'(x) + (b\,u')(x) + (c\,u)(x) = f(x)\,, \qquad 0 < x < 1\,, \tag{3.166}$$

with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$, where a, b and c are continuous functions on $[0, 1]$. Actually, such problems are used to describe the *diffusion* (the term $(a\,u')'$), the *advection* (the term $b\,u'$) and the *reaction* (the term cu) of a certain quantity $u(x)$. An interesting situation arises when a is small compared with b and c. To keep the experiment and the analysis simple, we consider the advection dominated boundary value problem:

$$-\varepsilon u'' + bu' = 0\,, \qquad 0 < x < 1\,, \tag{3.167}$$

with the boundary conditions $u(0) = 0$, $u(1) = 1$, where ε and b are two positive constants such that $\varepsilon/b \ll 1$. The analytical expresssion of the exact solution corresponds to:

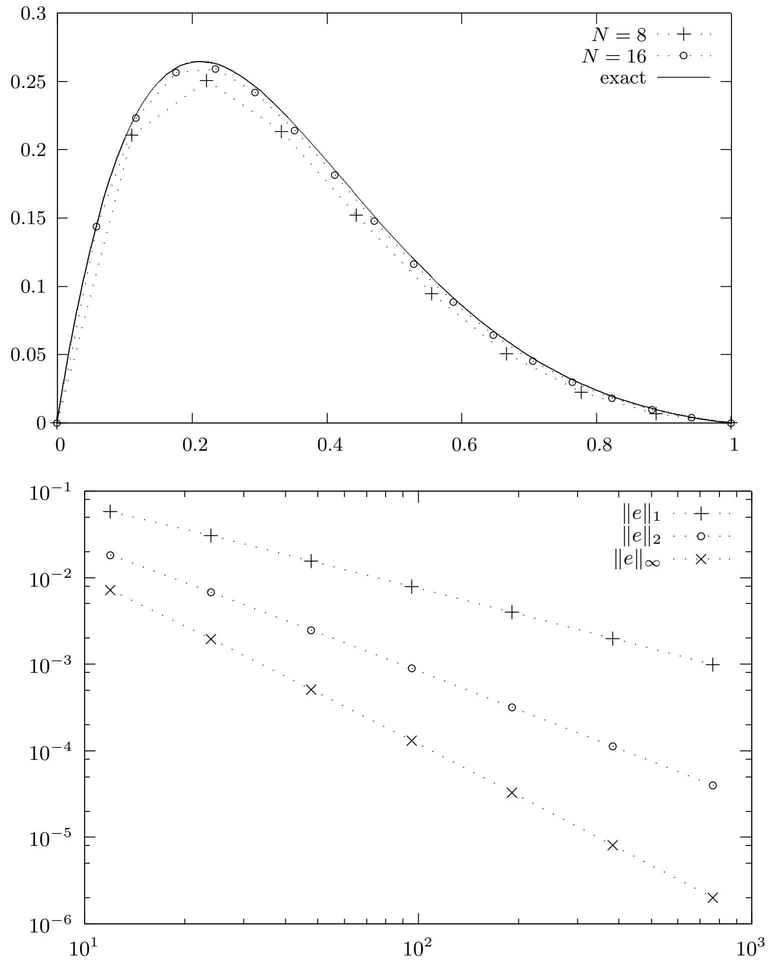

Fig. 3.3. *Numerical and exact solutions of the elliptic boundary value problem (3.164) (top) and convergence of the finite difference method with respect to the number of grid points.*

$$u(x) = \frac{\exp(b/\varepsilon\, x) - 1}{\exp(b/\varepsilon) - 1}\,, \tag{3.168}$$

We introduce the dimensionless *Péclet number* defined as:

$$\mathbb{P}e = \frac{|b|\, L}{2\varepsilon} \tag{3.169}$$

where L represents the size of the domain, which relates the relative importance of the advection term to the diffusion term. Since we assumed $\varepsilon/b \ll 1$, we observe that the solution is almost equal to zero in all the domain, except in a small neighborhood of the upper bound $x = 1$ where it reaches the value 1 exponentially. The width of the neighborhood is small, in order of ε/b: there is a *boundary layer* at the endpoint $x = 1$.

We computed a finite difference approximation of the solution using a second-order centered scheme, with a uniform distribution of the grid points $x_j = jh$, with the grid size $h = 1/N + 1$:

$$
\begin{gathered}
-\varepsilon \frac{uj + 1 - 2u_j + u_{j-1}}{h^2} + b\frac{u_{j+1} - u_{j-1}}{2h} = 0\,, \qquad 1 \leq j \leq N\,, \\
u_0 = 0\,, \qquad u_{N+1} = 1\,.
\end{gathered}
\tag{3.170}
$$

The result of the simulation is shown in Figure 3.4, where the numerical solution is compared to the exact solution, for a Péclet number $\mathbb{P}e = 50$, *i.e.* $b = 1$, $\varepsilon = 0.01$ and for $N = 10$ and $N = 30$. We shall notice the oscillations of the numerical solution in the vicinity of the boundary layer. One way of circumventing this problem consists in setting a sufficiently small grid size h, typically such that $\mathbb{P}e < 1/h$. However, this would rapidly reveal impractical when the Péclet number increases as the number of grid points becomes very large.

Another approach consists in replacing the advection term ε in the initial equation by $\varepsilon(1 + h\,\mathbb{P}e)$, thus solving the problem:

$$
-\varepsilon(1 + h\,\mathbb{P}e)\,u'' + b\,u' = 0\,, \qquad 0 < x < 1 \tag{3.171}
$$

with the boundary conditions $u(0) = 0$ and $u(1) = 1$. The *small* perturbation

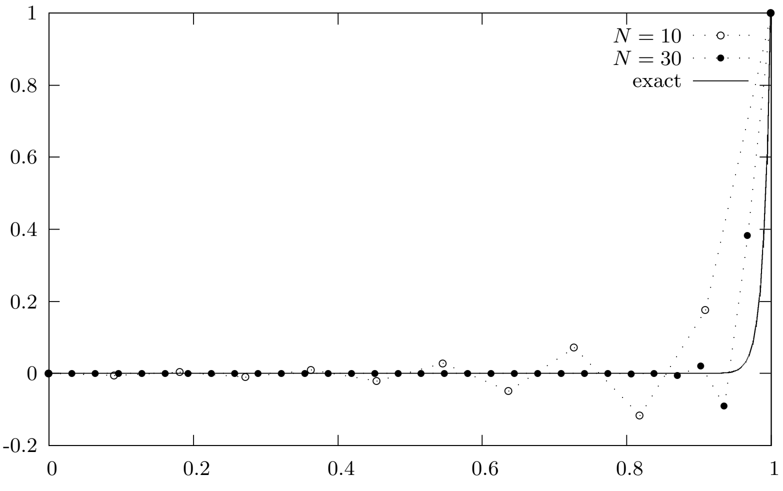

Fig. 3.4. *Exact solution (solid line) and finite difference solutions (dashed lines) of the advection-diffusion equation for $N = 10$ and $N = 30$ grid points and a Péclet number $\mathbb{P}e = 50$.*

$$-\varepsilon h \,\mathbb{P}e\, u'' = -\frac{|b|h}{2}\, u'' ,$$

is called the *numerical viscosity* and this technique is thus often called a *stabilization* method. In Figure 3.5, we observe the impact of the numerical viscosity on (almost) removing the oscillations of the numerical solution in the vicinity of the boundary layer.

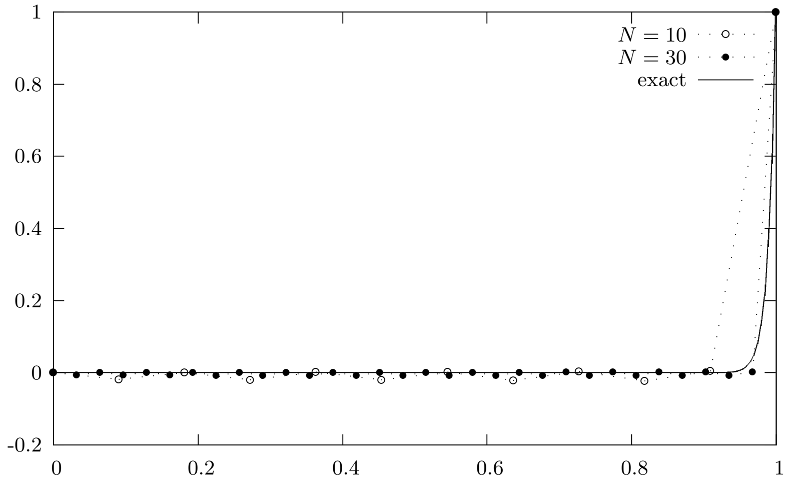

Fig. 3.5. *Exact solution (solid line) and finite difference solution (dashed line) of the stabilized advection-diffusion equation for $N = 10$ and $N = 30$ grid points and $\mathbb{P}e = 50$.*

3.7.2 Resolution of the 1d heat equation

From the analysis we carried out in Section 3.3, it is important to notice and to retain that an *unstable* numerical scheme is usually not useful for conducting numerical experiments, whatever the initial data considered.

For the first parabolic example, we considered the model problem:

$$\begin{cases} u_t = u_{xx} & \text{for } (x,t) \in]0,1[\times \mathbb{R}_+^* \,, \\ u(x,0) = u_0(x) & \text{for } x \in]0,1[\,, \\ u(0,t) = u(1,t) = 0 & \text{for } t \in \mathbb{R}_+^* \,. \end{cases} \tag{3.172}$$

with the initial data corresponding to $u_0(x) = \sin(2\pi x)$. We recall from Chapter 1 and the Fourier analysis that the exact solution is given by the equation:

$$u(x,t) = \exp(-4\pi^2 t)\, \sin(2\pi x)\,.$$

First, we solved the problem using the explicit scheme $(\mathcal{H}_1)$. To this end, we choose $N = 40$ grid points in space and set $\Delta t = 0.0001563$ and we compute the numerical solution at all times $0 \leq t^n \leq 0.02$. The analytical and numerical solutions at time $t = 0$, $t = 0.005$, $t = 0.01$ and $t = T$ are plotted in Figure 3.6, top. A piecewise linear interpolation is used between the grid points for the final time. In Figure 3.6, bottom, we have used the same grid points in space but a coarser time step $\Delta t = 0.0006875$. It is easy to see the impact of the time step Δt on the numerical solution. Here, the oscillations are mainly due to the accumulation of small round-off errors as the calculations

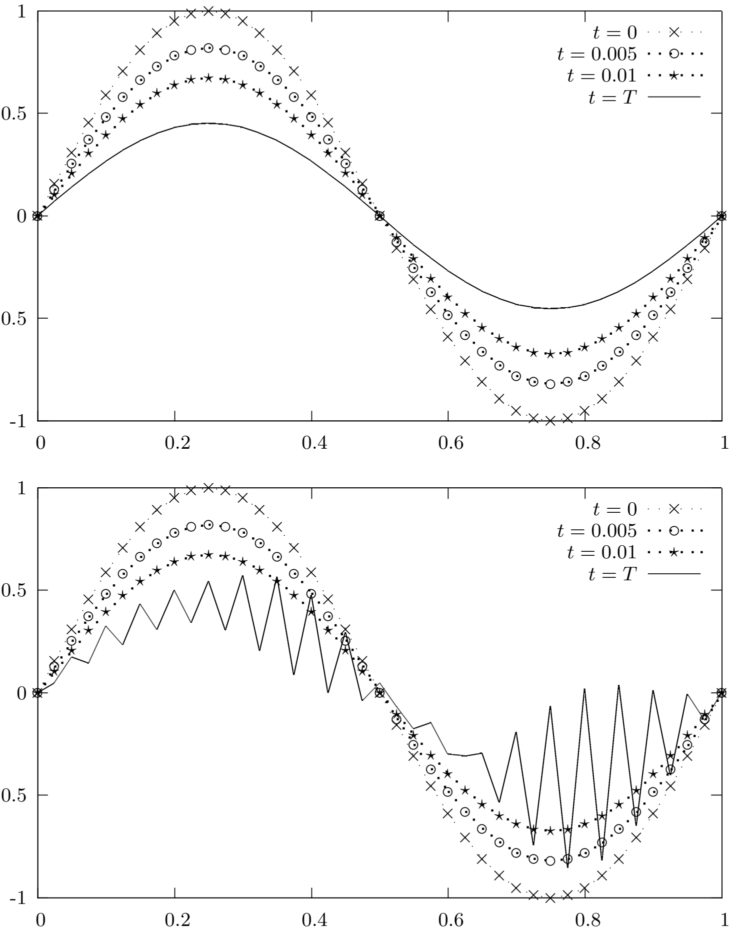

Fig. 3.6. *The finite difference method with the explicit scheme for solving the heat equation on a fixed grid* $N = 40$*. Top: the numerical solution at times* $t = 0$, $t = 0.005$, $t = 0.01$ *and* $t = T$ *with a time step* $\Delta t = 0.0001563$*. Bottom: the numerical solution for a time step* $\Delta t = 0.0006875$ *not preserving the stability condition.*

are performed on floats and not in exact arithmetic. To be correct, we shall mention here that the stability condition on the time step is in principle not strictly required, as the initial condition is a sine function. Nonetheless, this example is sufficient to show how the numerical computation can be heavily perturbated by a wrong or incorrect choice of the parameters.

Given the sinusoidal initial condition $u_0(x) = \sin(2\pi x)$, we computed the approximation errors at every grid points $e_j^n = u_j^n - u(x_j, t^n)$ on three different grids with 20, 30 and 50 points, respectively. The results are plotted in Figure 3.7, top. We have also computed the approximation errors using a θ-scheme, for three different values of the parameter θ. For this specific ini-

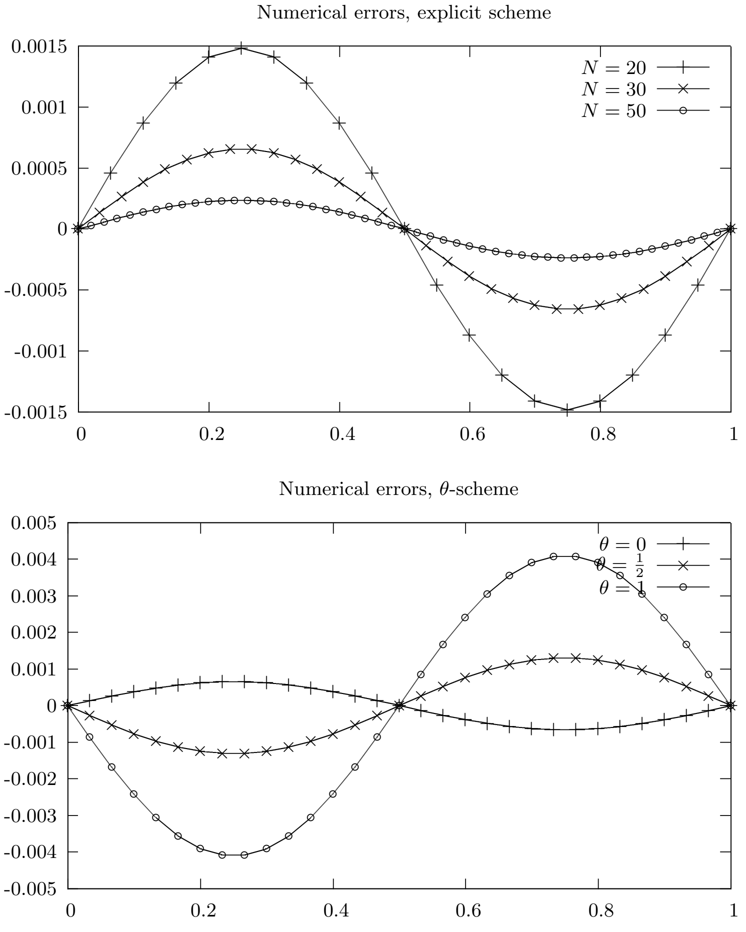

Fig. 3.7. *Comparison of the numerical errors. Top: explicit scheme with* 20, 30 *and* 50 *grid points. Bottom: θ-scheme on a fixed grid with* 30 *points, for the values of the parameter $\theta = 0$ (explicit), $\theta = \frac{1}{2}$ (Crank-Nicolson) and $\theta = 1$ (implicit).*

tial condition, we can observe that the approximation error for the explicit scheme is less important than the errors corresponding to the Crank-Nicolson or the implicit scheme (Figure 3.7, bottom). Notice however, that the Crank-Nicolson and implicit schemes are unconditionnally stable in L^2-norm, *i.e.,* without a restrictive condition on the time step.

It is well-known that solving the initial-value problem for the heat equation forward in time takes a discontinuous initial temperature u at time t_0 into a temperature which is instantly smooth as soon as the times $t > t_0$. This may not be physically possible, since there would then be information propagation at infinite speed, which is in contradiction with the *causality* principle. Therefore this is a theoretical property of the mathematical equation. This regularizing effect of the heat equation is illustrated on the Figure 3.8 for the initial condition $u_0(x) = \sin(\pi x) + \sin(5\pi x)$. We can easily observe the instantaneous decreasing of the oscillations in $\sin(5\pi x)$, before the fundamental mode starts to decrease itself. In other words, high frequencies are amortized first, thus having the positive effect of instantly removing all initialization errors. This interesting numerical property find numerous applications in image processing, mesh regularization, etc.

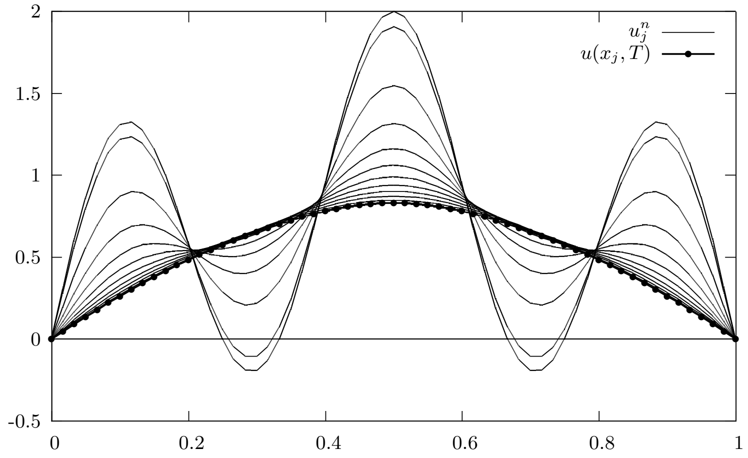

Fig. 3.8. *Regularization (smoothing) effect when solving the numerical heat equation at various time steps.*

3.7.3 Resolution of the 1d advection problem

We consider a simple example of a conservation law corresponding to the linear advection equation in a domain $\Omega = [0, L] \times [0, \infty[$:

$$u_t + c\,u_x = 0\,, \qquad x \in \Omega \times \mathbb{R}_+\,, \tag{3.173}$$

where $c(x,t) \in C^1(\Omega)$. This equation requires the specification of an initial condition and of boundary conditions to yield a well-posed problem, for which we know an analytical expression of the solution: $u(x,t) = u_0(x - ct)$. Here, we chose $c(x,t) = 1$ and $u(x,0) = u_0(x) = \max(1 - x^2, 0)$ in the translated domain $\Omega =]-10, 10[$ (here $L = 20$) and we considered periodic homogeneous boundary conditions: $u(-10,t) = u(10,t) = 0$. The domain Ω has been discretized in space using 500 grid points: $h = 0.04$.

We resolved the problem using two explicit schemes: the backward decentered (upwind) scheme (3.90) and the centered scheme (3.95). Figure 3.9 shows the numerical solutions obtained with these numerical schemes, both corresponding to the same CFL condition $\Delta t = 0.9\,h$. We shall observe that the numerical solution obtained with the centered scheme presents dramatic

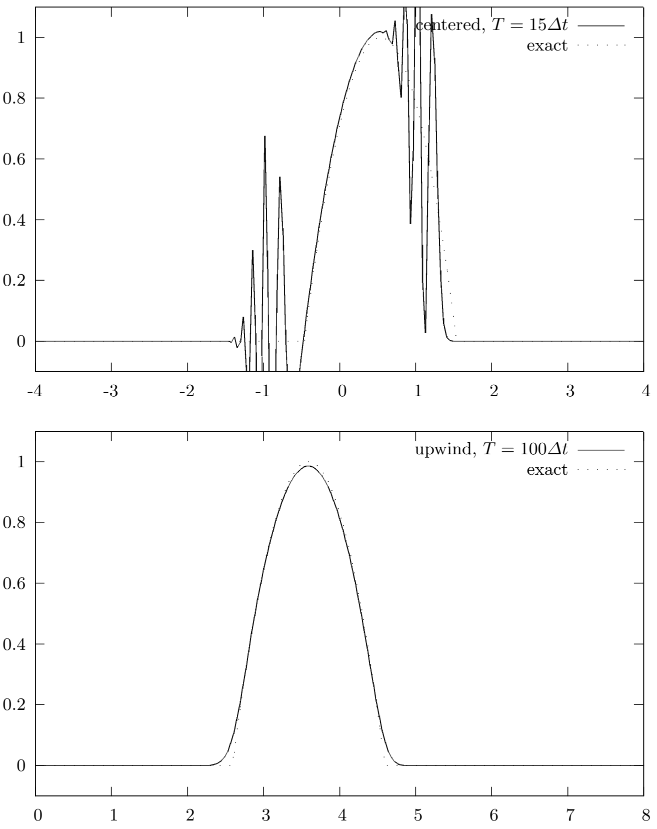

Fig. 3.9. *Numerical solutions obtained with two explicit schemes, for a CFL condition $\nu = 0.9$. top: the centered scheme (3.95) at time $t = 15\,\Delta t$. Bottom: the upwind scheme (3.90) at time $t = 100\Delta t$.*

oscillations after a few time steps, while the numerical solution obtained using the upwind scheme is close to the exact solution after 100 times steps. This confirms our theoretical analysis of the stability in L^2-norm of the upwind scheme under a CFL condition $\nu < 1$. This shows experimentally also that the centered scheme is unconditionally unstable in L^2-norm for solving the linear advection equation.

The nonlinear hyperbolic problems leads often to discontinuous solutions, like shock waves. Next, we analyze the behavior of the Lax-Wendroff and upwind numerical schemes described previously, when the initial data comprises such a discontinuity. To this end, we consider a model problem that consists in solving the advection equation:

$$u_t + cu_x = 0\,, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}_+\,, \tag{3.174}$$

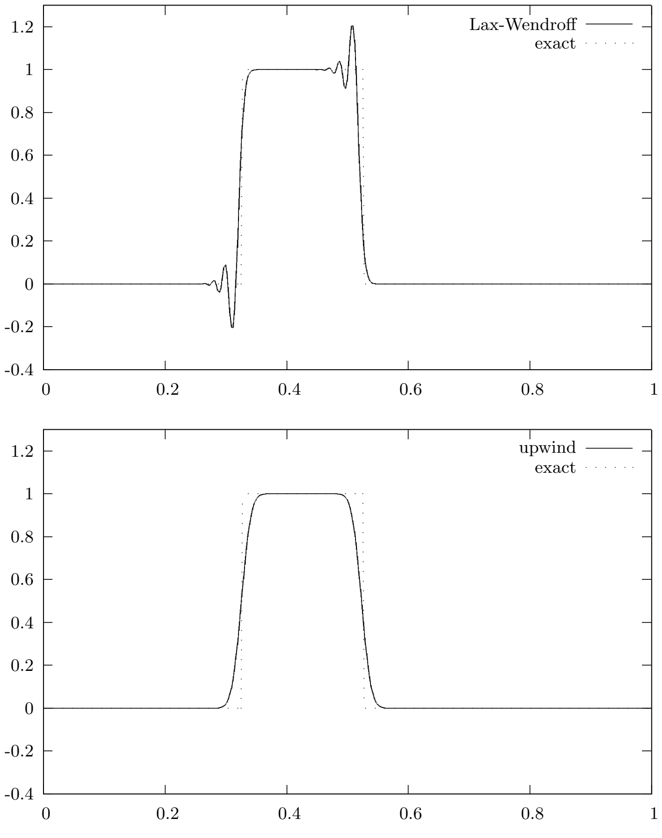

Fig. 3.10. *Numerical solutions obtained with two explicit schemes for a discontinuous solution: the explicit Lax-Wendroff scheme (left) and the upwind scheme (3.90) (right), for a CFL condition* $\nu = 0.5$.

with the initial and boundary conditions given by:

$$u(x,0) = u_0(x) = \begin{cases} 1 & \text{if } 0.2 < x \leq 0.4 \\ 0 & \text{otherwise} \end{cases} . \qquad u(x,0) = u(x+1,0)\,,\ \forall\, x \in \mathbb{R}\,. \tag{3.175}$$

In this specific case, the exact solution of this problem is still given by:

$$u(x,t) = u_0(x - c\,t)\,, \tag{3.176}$$

and represents a discontinuity (a square pulse) moving to the right.

In our example, we chose $c = 1$ and the CFL stability condition is satisfied by taking $\nu = \Delta t/h = 0.5$. Numerical results obtained using the explicit Lax-Wendroff and upwind schemes are presented in Figure 3.10 at time $T = 100\Delta t$ and are compared with the analytic solution (dashed line). We can clearly observe that, although the numerical solution moves at the correct speed, the upwind finite difference method preserves the amplitude of the initial shape, however the edges of the pulse are rounded off. The numerical solution is largely diffused and the discontinuity is spread out over roughly 20 grid points. The Lax-Wendroff scheme maintains the correct amplitude and width of the pulse but produces oscillations which follow behind the two discontinuities as the pulse evolves. Notice that reducing the grid size by a factor 2 would not change radically this situation and would only marginally improve the result, but not by the expected factor of 4, *i.e.,* as the truncation error is in $O(h^2)$. Indeed, the analysis of the error is valid only for smooth solutions, while the solution to this problem is clearly discontinuous. It could be shown that the error would be in $O(h^{2/3})$ for the Lax-Wendroff scheme.

Further reading

1. G. Allaire, *Analyse numérique et optimisation*, Editions de l'Ecole Polytechnique, (2005).
2. R. Haberman, *Elementary applied partial differential equations with Fourier series and boundary value problems*, 2nd edition, Prentice Hall, (1983).
3. R.J. LeVeque, *Finite difference methods for ordinary and partial differential equations*, SIAM, Philadelphia, (2007).
4. K.W. Morton and D. Mayers, *Numerical solution of partial differential equations. An introduction*, Cambridge University Press, 2nd edition, (2005).
5. A. Quarteroni and A. Vali, *Numerical approximation of partial differential equations*, Springer series in Computational Mathematics, Springer-Verlag, (1997).
6. A.A. Samarskii, *The theory of difference schemes*, Marcel Dekker Inc, New York, (2001).