Chapter 6

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.

Contents

6.1 Finite difference approximations

6.1.1 General principle

The principle of finite difference methods is close to the numerical schemes used to solve ordinary differential equations (cf. Appendix C). It consists in approximating the differential operator by replacing the derivatives in the equation using differential quotients. The domain is partitioned in space and in time and approximations of the solution are computed at the space or time points. 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 shall 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 $x \in \mathbb{R}$:

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

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.

6.1.2 Taylor series

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

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2} u''(x+h_1) \tag{6.1}$$

where h_1 is a number between 0 and h (*i.e.* $x + h_1$ is point of $]x, x+h[$). For the treatment of problems, it is convenient to retain only the first two terms of the previous expression:

$$u(x+h) = u(x) + hu'(x) + O(h^2)$$

where the term $O(h^2)$ indicates that the error of the approximation is proportional to h^2. From the equation (6.1), we deduce that there exists a constant $C > 0$, such that for $h > 0$ sufficienty small we have:

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

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 consistant at the first order. This approximation is known as the *forward difference* approximant of u'. More generally, we define an approximation at order p of the derivative.

Definition 6.1 *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:

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

Obviously, other approximations can be considered. In order to improve the accuracy of the approximation, we define a consistant approximation, called the *central difference* approximation, by taking the points $x - h$ and $x + h$ into account. Suppose that the function u is three times differentiable in the vicinity of x:

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

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:

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

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:

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

This defines a second order consistant approximation to u'.

Remark 6.1 *The order of the approximation is related to the regularity of the function u. If u is C^2 continuous, then the approximation is consistant at the order one only.*

6.1.3 Approximation of the second derivative

Lemma 6.1 *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:*

$$\left|\frac{u(x+h)-2u(x)+u(x-h)}{h^2} - u''(x)\right| \leq C\,h^2\,. \tag{6.2}$$

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

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

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi^+)$$
$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\xi^-)$$

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

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

where $\xi \in]x-h, x+h[$. Hence, we deduce the relation (6.2) with the constant

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

□

Remark 6.2 *Likewise, the error estimate depends on the regularity of the function u. If u is C^3 continuous, then the error is of order h only.*

6.2 Finite difference formulation for a one-dimensional 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:

$$\mathcal{D}\begin{cases} -u''(x) + c(x)u(x) = f(x)\,, & x \in]0,1[\,, \\ u(0) = \alpha\,,\ \ u(1) = \beta\,, \end{cases} \tag{6.3}$$

where c and f are two given functions, defined on $\bar{\Omega}$, $c \geq 0$.

6.2.1 Variational theory and approximation

Since Chapter 4, we know that if $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$, then the solution u to this problem exists. Furthermore, if $c = 0$, we have the explicit formulation of u as:

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

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 equation (6.3) is to partition the unit interval into a finite number of subintervals. Here, is a 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.

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 is given by $h = 1/(N+1)$. Typically, the spacing is aimed at becoming very small as the number of grid points will become very large. At the boundary of Ω, we have $x_0 = 0$ and $x_{N+1} = 1$. At each of these points, we are looking for numerical value of the solution, $u_j = u(x_j)$. We impose $u(x_0) = \alpha$ and $u(x_{N+1}) = \beta$ and we use the differential quotient introduced in the previous section to approximate the second order derivative of the equation (6.3).

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 $j \in \{1, \ldots, N\}$.

6.2.2 A finite difference scheme

Suppose functions c and f are at least such that $c \in C^0(\bar{\Omega})$ and $f \in C^0(\bar{\Omega})$. 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 problem (6.3). Introducing the approximation of the second order derivative by a differential quotient, we consider the following discrete problem:

$$\mathcal{D}_h \begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + c(x_j)u_j = f(x_j)\,, \qquad j \in \{1, \ldots, N\} \\ \qquad\qquad u_0 = \alpha\,,\ \ u_{N+1} = \beta\,, \end{cases} \tag{6.4}$$

The problem $\mathcal{D}$ has been discretized by a finite difference method based on a three-points centered scheme for the second-order derivative.

The problem (6.4) can be written in the matrix form as:

$$A_h u_h = b_h\,,$$

where A_h is the tridiagonal matrix defined as:

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

with

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

The question raised by this formulation is related to the existence of a solution. 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 6.1 *Suppose $c \geq 0$. Then, the matrix A_h is symmetric positive definite.*

Proof. We can observe that A_h is symmetric. Let consider a vector $v = (v_i)_{1 \leq i \leq N} \in \mathbb{R}^N$. Since $c \geq 0$, we have:

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

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

$$h^2 v^t\, A_h\, v = x_1^2 + (x_2 - x_1)^2 + \cdots + (x_{N-1} - x_N)^2 + x_N^2\,,$$

and thus $v^t\, A_h\, v \geq 0$. Moreover, if $v^t\, A_h\, v = 0$ then all terms $x_{i+1} - x_i = x_1 = x_N = 0$. Hence, we conclude that all $x_i = 0$ and the result follows. □

We can summarize the concept of finite differences for problem (6.3) in the following table:

	Theory (continuous)	Finite differences (discrete)
domain	$\Omega = [0,1]$	$I_N = \{0, \frac{1}{N+1}, \dots, 1\}$
unknown	$u : [0,1] \to \mathbb{R}$, $u \in C^2(\Omega)$	$u_h = (u_1, \dots, u_N) \in \mathbb{R}^N$
conditions	$u(0) = \alpha$, $u(1) = \beta$	$u_0 = \alpha$, $u_{N+1} = \beta$
equation	$-u'' + cu = f$	$-\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + c(x_j)u_j = f(x_j)$

6.2.3 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. We introduce a formal definition of the consistency that can be used for any partial differential equation defined on a domain Ω and denoted

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

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

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

For example, in the boundary value problem (6.3), the operator L is

$$(Lu)(x) = -u''(x) + c(x)u(x)\,,$$

and the problem can be written in the following form:
find $u \in C^2(\Omega)$ such that

$$(Lu)(x) = f(x)\,, \text{ for all } x \in \Omega\,. \tag{6.5}$$

We define the operator L_h by:

$$(L_h u)(x_j) = -\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + c(x_j)u(x_j)\,, \qquad j \in \{1, \ldots, N\}$$

and the discrete problem (6.4) can be formulated as follows:
find u such that

$$(L_h u)(x_j) = f(x_j)\,, \qquad \text{for all } j \in \{1, \ldots, N\}\,. \tag{6.6}$$

Definition 6.2 *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:*

$$(\varepsilon_h)_j = (L_h u)(x_j) - f(x_j)\,, \qquad \textit{for all } j \in \{1, \ldots, N\} \tag{6.7}$$

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 for the norm $\|\cdot\|$.

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

Lemma 6.2 *Suppose $u \in C^4(\Omega)$. Then, the numerical scheme (6.4) 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 $\xi_j \in]x_{j-1}, x_{j+1}[$. Hence, we have:

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

and the result follows. □

Remark 6.3 *Since the space dimension N is related to h by the relation $h(N+1)=1$, we have:*

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

The consistency error is a first step toward the analysis of the convergence error of the approximation method. However, it is not sufficient to analyze a numerical scheme.

Theorem 6.1 *Suppose $c \geq 0$ and that the solution of the problem $\mathcal{D}$ is of class $C^4(\Omega)$. Then, the finite difference scheme $\mathcal{D}_h$ is second-order convergent for the norm $\|\cdot\|_\infty$. Furthermore, if u and u_h are solutions of (6.5) and (6.6), we have the following estimate:*

$$\|u - u_h\|_\infty \leq \frac{h^2}{96} \sup_{x\in\Omega} |u^{(4)}(x)|\,.$$

Proof. Admitted here. □

6.3 Finite difference schemes for time-dependent problems

We consider a time-dependent, first-order boundary value problem posed in a bounded domain $\Omega =]0,1[$:

Find $u : [0,T] \times \bar{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) - \nu \dfrac{\partial^2 u}{\partial x^2}(t,x) = f(t,x)\,, & \forall t \in]0,T[\,, \quad \forall x \in]0,1[\\ u(0,x) = u_0(x)\,, & \forall x \in]0,1[\\ u(t,0) = u(t,1) = 0\,, & \forall t \in]0,T[\end{cases} \tag{6.8}$$

where $f(t,x)$ is a given source term and $\nu \geq 0$. This equation is the well-know *heat equation* that describes the distribution of heat in a given domain over time. It is the prototypical parabolic partial differential equation. The function $u(\cdot,\cdot)$, solution to this equation, describes the temperature at a given location x in time. The analysis of this equation has been pioneered by the French physicist J. Fourier (1768-1830) who invented influential methods for solving partial differential equations.

6.3.1 The continuous problem

We will show that this problem has a unique smooth solution that depends continuously on the data. The following estimate indicates the continuity, with respect to the data u_0 and f, of the solution u.

Lemma 6.3 *Suppose $u_0 \in L^2(\Omega)$ and $f \in L^2(]0,T[\times\Omega)$. If u is a sufficiently smooth solution of the problem (6.3.1), then we have the estimate:*

$$\sup_{t\in[0,T]} \|u(t,\cdot)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + \|f\|_{L^2(]0,T[\times\Omega)}\,.$$

Proof. Since the equation is linear, we have $u = v + w$, where v is solution of:

$$\begin{aligned} \frac{\partial v}{\partial t}(t,x) - \nu \frac{\partial^2 v}{\partial x^2}(t,x) = 0\,, & \qquad \forall x \in \Omega\,, \forall t \in]0,T[\,, \\ v(t,0) = v(t,1) = 0\,, & \qquad \forall t \in]0,T[\\ v(0,x) = u_0(x)\,, & \qquad \forall x \in \Omega\,. \end{aligned}$$

and similarly, w is solution of:

$$\begin{aligned} \frac{\partial w}{\partial t}(t,x) - \nu \frac{\partial^2 w}{\partial x^2}(t,x) = f(t,x)\,, \qquad & \forall x \in \Omega\,, \forall t \in]0,T[\,, \\ w(t,0) = w(t,1) = 0\,, \qquad & \forall t \in]0,T[\\ w(0,x) = 0\,, \qquad & \forall x \in \Omega\,. \end{aligned}$$

Hence, we can consider these two problems separately for v and w. Concerning the first problem, we multiply the equation by $v(x,t)$ and integrate it with respect to $x \in]0,1[$. Applying the chain rule and the integration by parts, and taking into account the homogenous boundary conditions, yields the following result:

$$\frac{1}{2}\frac{d}{dt}\left(\int_\Omega v^2(t,x)\,dx\right) + \nu \int_\Omega \left(\frac{\partial v}{\partial x}(t,x)\right)^2 dx = 0\,.$$

Since the second term is positive, we deduce:

$$\forall t \in]0,T[\,, \qquad \frac{d}{dt}\left(\int_\Omega v^2(t,x)\,dx\right) \le 0\,.$$

Now, we integrate with respect to the variable $t \in]0,s[$, $s \in [0,T]$ and thus we obtain:

$$\forall s \in [0,T]\,, \qquad \|v(s,\cdot)\|^2_{L^2(\Omega)} \le \|u_0\|^2_{L^2(\Omega)}\,.$$

We proceed accordingly for the term w, to obtain:

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\left(\int_\Omega w^2(t,x)\,dx\right) + \nu \int_\Omega \left(\frac{\partial w}{\partial x}(t,x)\right)^2 dx &= \int_\Omega (fw)(t,x)\,dx \\ &\le \|f(t,\cdot)\|_{L^2(\Omega)}\,\|w(t,\cdot)\|_{L^2(\Omega)}\,. \end{aligned}$$

Poincare's inequality leads to the estimate:

$$\|w(t,\cdot)\|_{L^2(\Omega)} \le C_p \left\|\frac{\partial w}{\partial t}(t,\cdot)\right\|_{L^2(\Omega)}\,,$$

and thus we have (noticing that $2ab \le (a^2+b^2)$):

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|w(t,\cdot)\|^2_{L^2(\Omega)} + \nu \left\|\frac{\partial w}{\partial t}(t,\cdot)\right\|^2_{L^2(\Omega)} &\le \|f(t,\cdot)\|_{L^2(\Omega)} \left\|\frac{\partial w}{\partial t}(t,\cdot)\right\|_{L^2(\Omega)} \\ &\le \frac{1}{2}\left(\|f(t,\cdot)\|^2_{L^2(\Omega)} + \left\|\frac{\partial w}{\partial x}(t,\cdot)\right\|^2_{L^2(\Omega)}\right)\,. \end{aligned}$$

Hence we can conclude that

$$\frac{d}{dt}\|w(t,\cdot)\|^2_{L^2(\Omega)} \le \frac{d}{dt}\|w(t,\cdot)\|^2_{L^2(\Omega)} + (2\nu - 1)\left\|\frac{\partial w}{\partial x}(t,\cdot)\right\|^2_{L^2(\Omega)} \le \|f(t,\cdot)\|^2_{L^2(\Omega)}\,.$$

We integrate on $]0,s[$ for $s \in [0,T]$ and taking into account the initial condition $w(x,0) = 0$, we obtain:

$$\forall s \in [0,T]\,, \qquad \|w(s,\cdot)\|^2_{L^2(\Omega)} \le \int_\Omega \|f(s,\cdot)\|^2_{L^2(\Omega)}\,ds \le \left(\|f\|_{L^2(\Omega)}\right)^2\,,$$

and, by combining this estimate with the estimate on v, the results follows. □

This results is important as it provides the uniqueness of a regular solution.

Corollary 6.1 *Suppose $u_0 \in L^2(\Omega)$ and $f \in L^2(]0,T[\times\Omega)$. Then, if the problem (6.3.1) has a regular solution, this solution is unique.*

Proof. Suppose $u_1(t,x)$ and $u_2(t,x)$ are to regular solutions of the problem (6.3.1). Then, if we denote their difference by $u = u_1 - u_2$, we obtain a problem similar to the initial one but where both the initial data u_0 and the source term are zero:

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) - \nu \dfrac{\partial^2 u}{\partial x^2}(t,x) = 0\,, & \forall (t,x) \in \mathbb{R}_+^* \times \Omega \\ u(0,x) = 0\,, & \forall x \in \Omega \\ u(t,x) = 0\,, & \forall (t,x) \in \mathbb{R}_+^* \times \partial\Omega \end{cases}$$

Then, using the previous estimate to this problem, we conclude that

$$\sup_{t\in[0,T]} \|u(t,\cdot)\|_{L^2(\Omega)} \leq 0\,,$$

and thus $u = u_1 - u_2 = 0$, the regular solutions are identical. □

Energy estimate

It is common to introduce the notion of *energy* of the solution u at time t as:

$$E(t) = \frac{1}{2}\int_\Omega u^2(t,x)\,dx = \frac{1}{2}\|u(t,\cdot)\|^2_{L^2(\Omega)} \tag{6.9}$$

This is not the physical energy of the system but rather a mathematical tool used to analyze the behavior of the solution. We shall see that using Lemma 6.3 and without the source term f, the energy is a non-increasing function of time. This result clearly indicates that this energy is controlled at any time t by the energy at the initial time $t = 0$, which is given. This important property of the heat equation must be preserved in the numerical resolution of the problem.

We introduce a preliminary result before giving an energy estimate.

Lemma 6.4 (Grönwall) *Let α, β be two real numbers and let u be a nonnegative real-valued function defined on $\mathbb{R}_+$ such that:*

$$u(t) \leq \alpha + \beta \int_0^t u(s)\,ds\,, \qquad t \in \mathbb{R}_+\,,$$

then $u(t) \leq \alpha \exp(\beta t)$.

Proof. Define $v(t) = \alpha + \beta \displaystyle\int_0^t u(s)\,ds$, $t \in \mathbb{R}_+$ so that $v(t) \geq u(t) \geq 0$. Taking the derivative and using the chain rule leads to $v'(t) = \beta v(t) \leq \beta v(t)$, and thus

$$v'(t) - \beta v(t) \leq 0\,, \qquad t \geq 0\,.$$

Multiplying by $\exp(-\beta t)$ and integrating between 0 and t yields

$$\int_0^t \exp(-\beta s) v'(s)\,ds - \beta v(s))\,ds \leq 0\,.$$

and after integrating by parts we have:

$$\int_\Omega \exp(-\beta s) v'(s)\,ds = \beta \int_0^t \exp(-\beta s) v(s)\,ds + \exp(-\beta t) v(t) - v(0)\,,$$

and finally the result follows: $v(t) \leq v(0)\exp(\beta t) = \alpha \exp(\beta t)$. □

Lemma 6.5 (Energy estimate) *Suppose $u_0 \in L^2(\Omega)$ and $f \in L^2(]0,T[\times\Omega)$. If u is a solution of problem (6.3.1) sufficiently smooth, then we have the following estimate:*

$$\int_\Omega u^2(t,x)\,dx \leq \exp(t)\left(\int_\Omega u_0^2(x)\,dx + \int_0^t \int_\Omega f^2(t,s)\,dxds\right).$$

Proof. We proceed as previously, multiply the equation (6.3.1) by u and integrate with respect to the variable x and then t to obtain:

$$\begin{aligned}\int_\Omega u^2(t,x)\,dx - \int_\Omega u_O^2(x)\,dx + 2\nu\int_0^t\int_\Omega \left(\frac{\partial u}{\partial x}(s,x)\right)^2 dxds \\ &= 2\int_\Omega f(s,x)u(s,x)\,dxds \\ &\le \int_0^t\int_\Omega f^2(s,x)\,dxds + \int_0^t\int_\Omega u^2(s,x)\,dxds\,.\end{aligned}$$

and we deduce:

$$\int_\Omega u^2(t,x)\,dx \le \int_\Omega u_0^2(x)\,dx + \int_0^T\int_\Omega f^2(s,x)\,dxds + \int_0^t\int_\Omega u^2(s,x)\,dxds\,.$$

Now, we invoke the Grönwall's lemma with

$$\alpha = \int_\Omega u_0^2(x)\,dx + \int_0^T\int_\Omega f^2(,x)\,dxds \qquad \text{and} \qquad \beta = 1\,,$$

to obtain:

$$\int_\Omega u^2(s,x)\,dx \le \exp(t)\left(\int_\Omega u_0^2(x)\,dx + \int_0^T\int_\Omega f^2(s,x)\,dxds\right), \qquad \forall T>0\,,$$

and the results follows. □

The estimates given by Lemmas 6.3 and 6.5 are often referred to as a *stability estimates*, since they express that the size of the solution can be bounded by the size of the initial data u_0 and f. A consequence of these results is that small perturbations of order ε of the initial data lead to small perturbations of the same order of the solution.

Corollary 6.2 *Suppose the initial data u_0 and f are slightly perturbated and replaced by new data $u_{0,\varepsilon} \in L^2(\Omega)$ and $f_\varepsilon \in L^2(]0,T[\times\Omega)$ such that:*

$$\|u_0 - u_{0,\varepsilon}\|_{L^2(\Omega)} \le \varepsilon\,, \qquad \text{and} \qquad \|f - f_\varepsilon\|_{L^2(]0,T[\times\Omega)} \le \varepsilon\,.$$

Then, if $u_\varepsilon(t,\cdot)$ denotes the new solution of the problem (6.3.1), we have the estimate:

$$\sup_{t\in[0,T]} \|u(t,\cdot) - u_\varepsilon(t,\cdot)\|_{L^2(\Omega)} \le \|u_0 - u_{0,\varepsilon}\|_{L^2(\Omega)} + \|f - f_\varepsilon\|_{L^2(]0,T[\times\Omega)} \le 2\varepsilon\,.$$

Proof. Since the problem (6.3.1) is linear, $u - u_\varepsilon$ solves the problem for the source term $f - f_\varepsilon$ and the initial data $u_0 - u_{0,\varepsilon}$. It is easy to see that for every $t \in [0,T]$ we have:

$$\|u(t,\cdot) - u_\varepsilon(t,\cdot)\|_{L^2(\Omega)} \le \|u_0 - u_{0,\varepsilon}\|_{L^2(\Omega)} + \|f - f_\varepsilon\|_{L^2(]0,T[\times\Omega)} \le 2\varepsilon\,.$$

and the results follows. □

6.3.2 Existence of a weak solution

In this section, we will briefly concentrate on the problem (6.3.1) with $u_0 \in L^2(\Omega)$ and $f \in L^2(]0,T[\times\Omega)$. Suppose the solution $u(t,\cdot) \in H^2(\Omega)$ for almost every $t \in]0,T[$. Let consider a test function $v \in H_0^1(\Omega)$ such that:

$$\int_\Omega \frac{\partial u}{\partial t}(t,x)v(x)\,dx - \nu\int_\Omega \frac{\partial^2 u}{\partial x^2}(t,x)v(x)\,dx = \int_\Omega f(t,x)v(x)\,dx\,.$$

Using the chain rule and integrating by parts, we obtain, for all $v \in H_0^1(\Omega)$:

$$\int_\Omega \frac{\partial u}{\partial t}(t,x)v(x)\,dx + \nu \int_\Omega \frac{\partial u}{\partial x}(t,x)\frac{dv}{dx}(x)\,dx = \int_\Omega f(t,x)v(x)\,dx\,.$$

We introduce a bilinear continuous and elliptic form $a(\cdot,\cdot)$ defined as:

$$a(w,v) = \langle w', v'\rangle_{L^2(\Omega)} = \int_\Omega w'(x)v'(x)\,dx\,,$$

that allows to write the previous problem (6.3.1) in a more compact form, for all $v \in H_0^1(\Omega)$:

$$\begin{cases} \dfrac{d}{dt}\langle u(t,\cdot), v\rangle_{L^2(\Omega)} + \nu a(u(t,\cdot), v) = \langle f(t,\cdot), v\rangle_{L^2(\Omega)}\,, \\ \langle u(0,\cdot), v\rangle_{L^2(\Omega)} = \langle u_0, v\rangle_{L^2(\Omega)}\,, \end{cases} \tag{6.10}$$

We give then the following result.

Theorem 6.2 *Suppose $u_0 \in L^2(\Omega)$ and $f \in L^2(]0,T[\times\Omega)$. Then, the problem (6.10) has a unique solution $u \in C^0([0,T]; L^2(\Omega)) \cap L^2(\{0,T\}; H_0^1(\Omega))$. Furthermore, we have the following energy estimate, for all $t \in [0,T]$:*

$$\|u(t,\cdot)\|_{L^2(\Omega)} \le \|u_0\|_{L^2(\Omega)} \exp(-\pi^2 t) + \int_0^t \|f(s,\cdot)\|_{L^2(\Omega)} \exp(-\pi^2(t-s))\,ds\,.$$

Proof. We refer the reader to [Allaire, 2005, Lucquin, 2004] for a proof of this result. □ This energy estimate shows that the solution of the problem (6.10) depends continuously on the data and thus that the problem is well-posed. Moreover, the energy space $C^0([0,T]; L^2(\Omega)) \cap L^2(\{0,T\}; H_0^1(\Omega))$ is the minimum regularity space in which the energy equalities are meaningfull.

6.3.3 An explicit scheme

To discretize the domain $[0,T]\times\bar{\Omega}$, we introduce equidistributed grid points corresponding to a spatial step size $h = 1/(N+1)$ and to a time step $\delta = 1/(M+1)$, where M, N are positive integers, and we define the nodes of a regular grid:

$$(t_n, x_j) = (n\delta, jh)\,, \qquad n \in \{0,\dots,M+1\}\,, \quad j \in \{0,\dots,N+1\}\,.$$

We denote as u_j^n the value of an approximate solution at point (t_n, x_j) and $u(t,x)$ the exact solution of problem (6.3.1). The initial data must also be discretized as:

$$u_j^0 = u_0(x_j)\,, \qquad \forall j \in \{0,\dots,N+1\}\,. \tag{6.11}$$

Finally, the homogeneous Dirichlet boundary conditions are discretized as:

$$u_0^n = u_{N+1}^n = 0\,, \qquad \forall n \in \{0,\dots,M+1\}\,. \tag{6.12}$$

The problem is then to find, at each time step, a vector $u_h \in \mathbb{R}^N$, such that its components are the values $(u_j^n)_{1\le j\le N}$.

Introducing the approximation of the second-order space derivative given by formula (6.2), and considering a forward difference approximation for the time derivative:

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^n}{\delta},$$

we obtain the following scheme coupled with the initial (6.11) and boundary conditions (6.12), for $n \in \{0, \dots, M\}$ and $j = \{1, \dots, N\}$:

$$\frac{u_j^{n+1} - u_j^n}{\delta} - \nu \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = f(t_n, x_j). \tag{6.13}$$

This scheme is called *explicit* because the values $(u_j^{n+1})_{1 \le j \le N}$ at time t_{n+1} are computed using the values on the previous time level t_n. Indeed, this system can be written in a vector form as:

$$\frac{U^{n+1} - U^n}{\delta} + A_h U^n = C^n, \qquad \forall n \in \{0, \dots, M\},$$

where the matrix A_h and the vector C^n are defined as:

$$A_h = \frac{\nu}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \quad \text{and} \quad C^n = \begin{pmatrix} f(t_1, x_1) \\ \vdots \\ \vdots \\ f(t_j, x_N) \end{pmatrix}. \tag{6.14}$$

and with the initial data:

$$U^0 = \begin{pmatrix} u_1^0 \\ \vdots \\ u_N^0 \end{pmatrix}.$$

To analyze the numerical scheme, we introduce a norm on $\mathbb{R}^N$:

$$\|u\|_p = \left(\sum_{j=1}^{N} h |u_j|^p \right)^{1/p}, \qquad \forall 1 \le p \le +\infty \tag{6.15}$$

where the limit case corresponds to $\|u\|_\infty = \max_{1 \le j \le N} |u_j|$. Notice that this norm depends on the step size $h = 1/N + 1$. Through the weight parameter h, the norm $\|u\|_p$ is identical to the $L^p(\Omega)$ norm for piecewise constant functions on each interval $[x_j, x_{j+1}[$ of the domain $\Omega = [0, 1]$.

Lemma 6.6 *The numerical scheme (6.13) is consistent and first-order accurate in time for the norm $\|\cdot\|_\infty$ and second-order accurate in space for the norm $\|\cdot\|_2$.*

Proof. Let consider a C^2 continuous in time and C^4 continuous in space function $u(t, x)$. We write:

$$\varepsilon_h(u)_j^n = \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{\delta} - \nu \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{h^2} - f(t_n, x_j).$$

If u is solution of the heat equation then we have:

$$\varepsilon_h(u)_j^n = A_j - \nu B_j,$$

where we posed

$$A_j = \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{\delta} - \frac{\partial u}{\partial t}(t_n, x_j)\,,$$
$$B_j = \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{h^2} - \frac{\partial^2 u}{\partial x^2}(t_n, x_j)$$

Using a Taylor expansion with respect to the time variable (x_j being fixed) we obtain, for $\tau \in]t_n, t_{n+1}[$:

$$u(t_{n+1}, x_j) = u(t_n, x_j) + \delta \frac{\partial u}{\partial t}(t_n, x_j) + \frac{\delta^2}{2}\frac{\partial^2 u}{\partial t^2}(\tau, x_j)\,,$$

and thus

$$A_i = \frac{\delta}{2}\frac{\partial^2 u}{\partial t^2}(\tau, x_j)\,,$$

and similarly for B_i we obtain for $\xi \in]x_j, x_{j+1}[$ (cf. Lemma (6.2)):

$$B_i = \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(t_n, \xi)\,,$$

We obtain easily the consistency and the order of accuracy of the explicit scheme from this formulas. □

Corollary 6.3 *If the ratio $\nu\delta/h^2 = 1/6$ is kept constant when δ and h tend towards zero, then the explicit scheme (6.13) is second-order accurate in time and fourth-order accurate in space.*

6.3.4 Other schemes

Changing the approximation of the time derivative for a backward difference would have resulted in the following *implicit* scheme:

$$\frac{u_j^n - u_j^{n-1}}{\delta} - \nu\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = f(t_n, x_j)\,. \tag{6.16}$$

that requires solving a system of linear equations to compute the values $(u_j^n)_{1\le j\le N}$ at time t_n using the values of the previous time level t_{n-1}. The scheme can be written in vector form as:

$$\frac{U^n - U^{n-1}}{\delta} + A_h U^n = C^n\,, \qquad \forall n \in \{1, \ldots, M+1\}\,, \tag{6.17}$$

where the matrix A_h and the vectors C^n and U^0 are identical to the corresponding terms in the explicit scheme (6.13). However, computing the values u_j^n with respect to the values u_j^{n-1} requires finding the inverse of the tridiagonal matrix A_h.

Proposition 6.2 *The matrix A_h is positive definite and thus is invertible.*

Lemma 6.7 *The numerical scheme (6.17) is consistent and first-order accurate in time for the norm $\|\cdot\|_\infty$ and second-order accurate in space for the norm $\|\cdot\|_2$.*

Proof. Left as an exercise. □

Using a convex combination of the explicit and implicit schemes, we define for $0 \le \theta \le 1$, the so-called θ-scheme:

$$\frac{u_j^n - u_j^{n-1}}{\delta} - \theta\nu\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} - (1-\theta)\nu\frac{u_{j+1}^{n-1} - 2u_j^{n-1} + u_{j-1}^{n-1}}{h^2} = \theta f(t_n, x_j) + (1-\theta) f(t_{n-1}, x_j)\,, \tag{6.18}$$

that gives the explicit (resp. implicit) scheme if $\theta = 0$ (resp. $\theta = 1$). Notice that this scheme is implicit for $\theta \neq 0$. For the value $\theta = 1/2$, the scheme is called the *Crank-Nicholson* scheme.

All these schemes are multi-level schemes (here two levels) as they involve two time indices u^n and u^{n-1} so that the matrix of the system is tridiagonal, *i.e.* has non-zero elements only on the diagonal and the positions immediately to the left and to the right of the diagonal. Here are a few other examples of multi-level schemes:

Richardson scheme:
$$\frac{u_j^{n+1}-u_j^{n-1}}{2\delta}-\nu\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{h^2}=f(t_n,x_j)$$

DuFort-Frankel scheme:
$$\frac{u_j^{n+1}-u_j^{n-1}}{2\delta}+\nu\frac{u_j^{n+1}-u_{j+1}^n-u_{j-1}^n+u_j^{n-1}}{h^2}=f(t_n,x_j)$$

Gear scheme:
$$\frac{3u_j^{n+1}-4u_j^n+u_j^{n-1}}{2\delta}-\nu\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{h^2}=f(t_n,x_j)\,.$$

We can then consider a more general form for such systems:

$$B_lU^{n+l}+B_{l-1}U^{n+l-1}+\cdots+B_0U^n+\cdots+B_{-m}U^{n-m}=C^n\,,$$

for $n\geq m$, $l,m\geq 0$, $l+m\geq 1$, B_l invertible and $U^0,\ldots,U^{l+m-1}$ given. Such a scheme involves $l+m$ levels. If the matrix B_l is diagonal, the scheme is explicit and implicit otherwise.

6.3.5 Stability and Fourier analysis

Roughly speaking, the instability consists in the emergence of unbounded oscillations in the numerical solution.

Definition 6.3 *A finite difference scheme is said to be* stable *for the norm $\|\cdot\|_p$ defined by (6.15), if there exists two constants $C_1>0$ and $C_2>0$, independent of h and δ, such that when h and δ tend towards zero:*
$$\|u\|_p\leq C_1\|u_0\|+C_2\|f\|\,,\qquad\forall n\geq 0 \tag{6.19}$$
whatever the initial data u_0 and the source term f.

Remark 6.4 *Suppose $f\in L^\infty(]0,T[\times\Omega)$ and if we consider the norm $\|\cdot\|_p$ on $\mathbb{R}^N$, then the previous stability estimate is exactly the discrete analoge of the estimate provided by Lemma 6.3 (with $C_1=1$ and $C_2=\sqrt{T}$).*

Remark 6.5 *Since all norms are equivalent in $\mathbb{R}^N$ it would be tempting to conclude that the stability with respect to a norm implies the stability with respect to all other norms. Unfortunately, this is not true because in the definition, the bound is uniform with respect to h but the norms $\|\cdot\|_p$ depend on h.*