The Finite Difference Method

Brigitte Lucquin, Olivier Pironneau

December 4, 1999

Premia 2

This is a topic of

Finite Difference Methods

Contents

1 Summary

We introduce here the finite difference method for approximating the three types of Partial Differential Equations: elliptic, parabolic and hyperbolic. We present the principle of the method on the example of a simple elliptic equation, then we shall focus our study on parabolic equations. The generalization to several space dimensions is described in the case of the heat equation. If the spatial mesh is not uniform, it might be more advisable to use the finite element method, or the finite volume method, which we shall briefly introduce here also. Important topics like stability convergence consistency and the maximum principle are presented

At the end of the chapter we consider the special case of the Black and Scholes equation for European and American options.

2 Finite differences

2.1 Introduction

Let us go back to the case of a rod heated at both ends, which we studied in the Introduction to this book. The temperature φ of this rod, taken to be a line segment of unit length, solves the variational problem:

$$-\frac{d^2\varphi}{dx^2}(x) + c(x)\varphi(x) = f(x), \quad 0 < x < 1, \tag{1}$$

$$\varphi(0) = g(0), \quad \varphi(1) = g(1). \tag{2}$$

If the function c takes positive values, this problem admits a unique solution. If c is zero, the exact solution to (1)-(2) is given by

$$\varphi(x) = \int_0^1 G(x,y)f(y)dy + g(0) + x(g(1) - g(0)), \tag{3}$$

where G is the Green function, defined by

$$G(x,y) = \{\, (1-x)y, \text{ if } 0 \le y \le x, (1-y)x, \text{ if } x \le y \le 1. \tag{4}$$

Apart from this particular case, and in any case when the problem is set in dimension $d \ge 1$, the solution cannot be computed explicitly, and a discretization is required, so that we can give as accurate an approximation as possible.

Contrary to the finite element method, the finite difference method consists in *approximating the derivation operator* by a discrete operator. This approach is easily understood if we notice that, for small h, we have (for instance):

$$\frac{\partial\varphi}{\partial x_i}(x_1, \cdots, x_d) \simeq \frac{1}{h}[\varphi(x_1, \cdots, x_i + h, \cdots, x_d) - \varphi(x_1, \cdots, x_d)]. \tag{5}$$

This remark suggests that we "replace" all continuous derivative operators by difference quotients (whence the name "difference", "finite"coming from the fact that the parameter h, though chosen arbitrarily small, has a fixed nonzero value). The *a posteriori* justification of this approximation results from using simple Taylor formulas.

We shall now study some examples of discrete operators obtained this in way, then deduce a approximation to the one dimensional problem (1)-(2).

2.2 Examples of difference operators

We shall work in dimension d and, to simplify notations, we write φ for $\varphi(x_1, \cdots, x_d)$ and $\varphi(x_i + h)$ for $\varphi(x_1, \cdots, x_i + h, \cdots, x_d)$. Let us introduce the following linear operators

$$D_i^+\varphi = \frac{1}{h}(\varphi(x_i + h) - \varphi), \tag{6}$$

$$D_i^-\varphi = \frac{1}{h}(\varphi - \varphi(x_i - h)), \tag{7}$$

$$D_i^o\varphi = \frac{1}{h}(\varphi(x_i + \frac{h}{2}) - \varphi(x_i - \frac{h}{2})). \tag{8}$$

The operator D_i^o is called a "*centered operator*" in direction i, whereas the other two are "non-centered": forward for the first one, backward for the second one.

The very definition of the derivative shows that $D_i^+\varphi$ and $D_i^-\varphi$ tend towards $\frac{\partial\varphi}{\partial x_i}$ when h goes to 0. For the centered operator, this stems from taking the difference between the following two Taylor expansions:

$$\varphi(x_i + \frac{h}{2}) = \varphi(x_i) + \frac{h}{2}\frac{\partial\varphi}{\partial x_i}(x_i + \theta_1(h)), \quad \lim_{h\to 0}\theta_1(h) = 0, \tag{9}$$

$$\varphi(x_i - \frac{h}{2}) = \varphi(x_i) - \frac{h}{2}\frac{\partial\varphi}{\partial x_i}(x_i + \theta_2(h)), \quad \lim_{h\to 0}\theta_2(h) = 0. \tag{10}$$

We say these operators are *consistent* approximations to $\frac{\partial\varphi}{\partial x_i}$. If furthermore the error $|D_i\varphi - \frac{\partial\varphi}{\partial x_i}|$ thus committed is bounded, up to a constant, by h^p, the approximation is said to be *consistent of order p*.

approximations ? consistent

With a view towards solving our initial problem, we shall now define an approximation for the second derivatives.

Proposition 1.1 *If φ is 4 times continuously differentiable in the interval $[x_i - h, x_i + h]$,*

$$D_i^oD_i^o\varphi = D_i^+D_i^-\varphi = D_i^-D_i^+\varphi = \frac{1}{h^2}[\varphi(x_i + h) - 2\varphi + \varphi(x_i - h)] \tag{11}$$

is a consistent, second order, approximation to $\frac{\partial^2\varphi}{\partial x_i^2}$.

Proof First of all we have

$$D_i^o(D_i^o\varphi) = \tfrac{1}{h}(D_i^o\varphi|_{x_i+\frac{h}{2}} - D_i^o\varphi|_{x_i-\frac{h}{2}}) = \tag{12}$$

$$\tfrac{1}{h}(\tfrac{\varphi(x_i+h)-\varphi}{h} - \tfrac{\varphi-\varphi(x_i-h)}{h}) = \tfrac{1}{h^2}[\varphi(x_i+h) - 2\varphi + \varphi(x_i-h)]. \tag{13}$$

Next, an analogous computation shows that $D_i^oD_i^o\varphi = D_i^+D_i^-\varphi = D_i^-D_i^+\varphi$, which proves (11).

Now, if φ is of class C^4 on $[x_i-h, x_i+h]$, we can write the following Taylor expansions

$$\varphi(x_i+h) = \varphi + h\tfrac{\partial\varphi}{\partial x_i} + \tfrac{h^2}{2}\tfrac{\partial^2\varphi}{\partial x_i^2} + \tfrac{h^3}{6}\tfrac{\partial^3\varphi}{\partial x_i^3} + \tfrac{h^4}{24}\tfrac{\partial^4\varphi}{\partial x_i^4}(\xi^+), \quad \xi^+ \in]x_i, x_i+h[\tag{14}$$

$$\varphi(x_i-h) = \varphi - h\tfrac{\partial\varphi}{\partial x_i} + \tfrac{h^2}{2}\tfrac{\partial^2\varphi}{\partial x_i^2} - \tfrac{h^3}{6}\tfrac{\partial^3\varphi}{\partial x_i^3} + \tfrac{h^4}{24}\tfrac{\partial^4\varphi}{\partial x_i^4}(\xi^-), \quad \xi^- \in]x_i-h, x_i[\tag{15}$$

and adding them gives

$$D_i^oD_i^o\varphi = \frac{\partial^2\varphi}{\partial x_i^2} + \frac{h^2}{24}[\frac{\partial^4\varphi}{\partial x_i^4}(\xi^+) + \frac{\partial^4\varphi}{\partial x_i^4}(\xi^-)]. \tag{16}$$

By the mean value theorem, we deduce the existence of a number $\xi \in]x_i-h, x_i+h[$ such that

$$D_i^oD_i^o\varphi - \frac{\partial^2\varphi}{\partial x_i^2} = \frac{h^2}{12}\frac{\partial^4\varphi}{\partial x_i^4}(\xi), \tag{17}$$

which shows that the consistency error $|D_i^oD_i^o\varphi - \frac{\partial^2\varphi}{\partial x_i^2}|$ is bounded by Ch^2, with $12\,C = \sup_{\xi\in]x_i-h,x_i+h[} |\frac{\partial^4\varphi}{\partial x_i^4}(\xi)|$. So the approximation is consistent of order 2.

By combining in different ways the operators D^+, D^- and D^o, it is possible to construct approximations to a partial derivative of any arbitrary order; some are better than others, meaning that their consistency order is higher.

2.3 Finite difference approximation

We partition the segment $[0,1]$ into $N+1$ intervals of length $h = \delta x = 1/(N+1)$, and we define the $N+2$ *subdivision points*, or *nodes*, of this regular mesh by $x_i = ih$, $i \in \{0, ..., N+1\}$.

The discrete problem will be to find an approximation ψ_i to $\varphi(x_i)$ at each internal (*i.e.* $x_i, 1 \le i \le N$) node of the mesh, since, by virtue of (2), the solution is known at the ends $x_0 = 0$ and $x_{N+1} = 1$ of the interval. These different values ψ_i are solutions of the discrete problem

$$-\tfrac{1}{h^2}[\psi_{i+1} - 2\psi_i + \psi_{i-1}] + c(x_i)\psi_i = f(x_i), \quad 1 \le i \le N, \tag{18}$$

$$\psi_0 = g(0), \quad \psi_{N+1} = g(1), \tag{19}$$

which we call a *finite difference scheme* for problem (1)-(2). If we denote by Ψ_h the unknown vector with entries $(\psi_1, ..., \psi_N)^T$, this problem (18)-(19) can be written in matrix form

$$A_h \Psi_h = b_h, \tag{20}$$

where the symmetric matrix A_h and the right hand side b_h are given by

$$A_h = \frac{1}{h^2} \begin{bmatrix} 2 + c(x_1)h^2 & -1 & & & 0 \\ -1 & 2 + c(x_2)h^2 & -1 & & \\ & & \dots & & \\ & & -1 & 2 + c(x_{N-1})h^2 & -1 \\ 0 & & & -1 & 2 + c(x_N)h^2 \end{bmatrix},$$

$$b_h = \begin{pmatrix} f(x_1) + \frac{1}{h^2} g(0) \\ f(x_2) \\ \dots \\ f(x_{N-1}) \\ f(x_N) + \frac{1}{h^2} g(1) \end{pmatrix}.$$

More generally, we shall adopt the following definition:

Definition 1.2 Let $L\varphi = 0$ be a partial differential equation and let $L_h \Psi_h = 0$ be a finite difference scheme of the above kind for approximating this problem; we shall call *consistency error* of the scheme the vector ε_h defined by

$$\varepsilon_h = L_h \varphi_h, \tag{21}$$

where φ_h is the projection on the mesh of the *exact solution* φ of the continuous problem, *i.e.* if the mesh is formed by the N points x_i, then φ_h is the vector with entries $(\varphi(x_1), ..., \varphi(x_N))^T$. The scheme is said to be *consistent* if this vector tends to 0 with h. This requires that we have defined a vector norm on R^N. Let us define, for example, the norm $||.||_\infty$ defined by $||X||_\infty = \sup_{i=1}^{N} x_i$, $X = (x_1, ..., x_N)^T$. We shall say that the *scheme is of order p* in the l^∞ norm if there is a real positive constant C such that

$$||\varepsilon_h||_\infty \leq C h^p. \tag{22}$$

According to proposition 1.1, we deduce the following result:

Corollary 1.3 *If the exact solution of problem (1)-(2) is of class C^4 on $[0, 1]$, the scheme (18)-(19) is consistent of order 2.*

Remark Still using only the 3 points x_{i-1}, x_i and x_{i+1}, it is possible to construct fourth order approximations to the first and second derivatives. Indeed, show that, if φ is of class C^6, then

$$\varphi'(x_{i+1}) + 4\varphi'(x_i) + \varphi'(x_{i-1}) = \frac{3}{h} [\varphi(x_{i+1}) - \varphi(x_{i-1})] + 0(h^4), \tag{23}$$

$$\varphi"(x_{i+1}) + 10\varphi"(x_i) + \varphi"(x_{i-1}) = \frac{12}{h^2} [\varphi(x_{i+1}) - 2\varphi(x_i) + \varphi(x_{i-1})] + 0(h^4) \tag{24}$$

By linearly combining the equations $-\varphi"(x_k) = f_k$, $k \in \{i+1, i, i-1\}$, this enables us, for instance, to obtain the following scheme for the Laplacian in one dimension,

$$-\frac{12}{h^2}[\psi_{i+1} - 2\psi_i + \psi_{i-1}] = f(x_{i+1}) + 10f(x_i) + f(x_{i-1}). \tag{25}$$

Show that this scheme is fourth order in the l^∞ norm

Still regarding problem (1)-(2), two questions then arise:

- (Q1): does the discrete problem admit *a unique solution?*
- (Q2): is the method *convergent*, that is does it hold that

$$||\Psi_h - \varphi_h||_\infty \to 0, \quad \text{when } h \to 0? \tag{26}$$

It is easy to give an affirmative answer to the first question, since a simple computation shows that the matrix A_h is positive definite; indeed, if $X = (x_1, ..., x_N) \in R^N$, we have

$$X^T A_h X = x_1^2 + (x_2 - x_1)^2 + ... + (x_N - x_{N-1})^2 + x_N^2 + h^2 \sum_{i=1}^{N} c(x_i) x_i^2, \tag{27}$$

and this quantity is positive, because of the positiveness assumption on c, and can only be zero if all the x_i are zero.

The answer to the second question is also positive. This stems from the *consistency* of the scheme and from *stability* results due to the monotonicity of the matrix A_h. We shall not prove these results here, as we reserve this stability study for the case of *parabolic and hyperbolic* problems, which are the main goal of this Chapter. The results for the case of elliptic operators, such results are proved in Ciarlet (1988), to which we refer the reader .

Remark 1.4 Since the matrix A_h is symmetric and positive definite, several methods are possible for actually solving system (20: Choleski's method, the Gauss-Seidel or conjugate gradient algorithms. Because the matrix is diagonally dominant, Jacobi's method also converges.

Remark 1.5 The above study can be generalized to elliptic operators in *dimension larger than one.* For example, in the case of the Laplacian with Dirichlet boundary conditions on a rectangular domain in the plane,

$$-\Delta\varphi = f \text{ in } \Omega =]0, L_1[\times]0, L_2[, \tag{28}$$

$$\varphi = g \text{ on } \Gamma = \partial\Omega, \tag{29}$$

the mesh is made up of small elementary rectangles of length h_i along each axis x_i, $i \in \{1, 2\}$, and, in 2 dimensions, the scheme can be written, for the internal node of the mesh with index (i, j) :

$$-\frac{1}{(h_1)^2}[\psi_{i+1,j} - 2\psi_{i,j} + \psi_{i-1,j}] - \frac{1}{(h_2)^2}[\psi_{i,j+1} - 2\psi_{i,j} + \psi_{i,j-1}] = f(ih_1, jh_2), \tag{30}$$

where $\psi_{ij} \simeq \varphi(ih_1, jh_2)$. The matrix of the resulting linear system is again symmetric and positive definite; it is pentadiagonal and block-tridiagonal, with each diagonal block itself a tridiagonal matrix. We shall have the opportunity to come back to this point when we treat the heat equation in 2 dimensions. This scheme is called a "5 points scheme" for the Laplacian (*cf.* figure VII.1).

the 5 points scheme for the Laplacian

Remark 1.6 A last remark to conclude: how can we handle *Neumann boundary conditions?* Mathematically, the answer is much less clear than it was for the finite element method, for which this type of boundary conditions was taken into account naturally by the variational formulation. It can be useful, for more complicated operators than the mere Laplacian, to combine this variational formulation to "ad-hoc" quadrature formulas, so as to find the "right scheme" near the boundary [Lucquin-Pironneau (1997)].

For the one-dimensional problem

$$-\varphi"(x) = f(x), \quad 0 < x < 1, \tag{31}$$

$$-\varphi'(0) = g(0), \quad \varphi'(1) = g(1), \tag{32}$$

we can for instance propose, keeping the same notation as above, the following approximation of the boundary condition

$$\psi_0 - \psi_1 = hg(0), \quad \psi_N - \psi_{N-1} = hg(1). \tag{33}$$

However, the accuracy of the global scheme is only h near the boundary.

By a clever linear combination (determined thanks to Taylor's formula) involving additional interior nodes, it is possible to improve the accuracy of the boundary condition approximation.

For large problem in several space dimensions it is not advisable to use a direct method like Choleski's au Gauss'; one of the best iterative scheme available is described in Appendix A: conjugate gradient with BPX preconditionning;

3 Finite difference schemes for linear evolution problems

The main goal of this Chapter is to define an approximation for *parabolic and hyperbolic* problems, then study its properties. In this type of equations, one of the variables, called the "time variable", is singled out, contrary to elliptic equations. The other variables will be called space variables; they lie in the whole of R^d or in a domain in R^d. Such problems are called *evolution problems* in time, as the solution at time $t \geq 0$ is determined from values at time $t = 0$, which we call "initial conditions".

We shall define a numerical approximation, using finite differences for the time variable, for these evolution problems, and we shall assume that their space dependence is limited to linear partial differential operators.

3.1 Introduction

Let us consider the following Cauchy problem

$$\frac{\partial \varphi}{\partial t}(t) = A(\varphi(t)), \quad 0 \leq t \leq T, \tag{34}$$

$$\varphi(t=0) = \varphi^0, \tag{35}$$

where A is a differential operator, assumed linear and independent of the t variable; for example, A can be the Laplacian, possibly with boundary conditions (if it does not act on the whole space). It is clear that a possible solution φ of problem (34)-(35) also depends on a space variable x that we have deliberately left out of the equations, on the one hand to simplify the notation, but mostly to emphasize the role of the time variable.

Let us assume that this problem has a classical solution $(t \to \varphi(t))$ in some function space, and let us then denote by $S(t)$ the operator defined by

$$S(t)\varphi^0 = \varphi(t), \tag{36}$$

where $\varphi(t)$ is the solution at time t of problem (34)-(35).

The approximation using *finite differences in time* of this problem consists in partitioning the interval under consideration $[0, T]$ into M subintervals of length $k = \Delta t = T/M$, then, given an approximation ψ^n of $\varphi(t^n)$, $t^n = n\delta t$, in defining an approximation ψ^{n+1} of the exact solution

$$\varphi(t^{n+1}) = S(k)\varphi(t^n) \tag{37}$$

of problem (34)-(35) at the next time step as

$$\psi^{n+1} = G(k)\psi^n, \tag{38}$$

where $G(k)$ is the "discretized in space counterpart" of the operator $S(k)$, in particular meaning that it depends on a space discretization parameter denoted by $h = \delta x$, assumed to be as small as we wish. In the same way, what we denote by ψ^n is actually a vector, each of whose entries corresponds to its value at a node of the mesh, but, as for the continuous problem, we have not displayed this spatial dependence explicitly.

Naturally, this iterative construction procedure requires the knowledge of the approximate solution at the initial time, and we shall simply define it as

$$\psi^0 = \varphi(t=0) = \varphi^0. \tag{39}$$

What are the properties of such a scheme? How can we measure the error? Building upon the definition of consistency given in the previous Section, we shall first answer the second question.

3.2 Consistency of scheme

By analogy with the notation used in definition 1.2, we denote by L the continuous operator

$$L = \frac{\partial}{\partial t} - A, \tag{40}$$

and we define the discrete operator L^k ($k = \delta t$) on sequences $\Psi^k = (\psi^n)_{n\geq 0}$, $\psi^0 = \varphi^0$ in the following way: $L^k\Psi^k$ is the sequence defined by

$$L^k\Psi^k = \left((L^k\Psi^k)^n \right)_{n\geq 1}, \quad (L^k\Psi^k)^n = \frac{\psi^n - G(k)\psi^{n-1}}{k}, \tag{41}$$

since according to (38), we have:

$$\frac{\psi^n - \psi^{n-1}}{k} - \frac{G(k)\psi^{n-1} - \psi^{n-1}}{k} = 0. \tag{42}$$

Let us note that this operator also depends *on the space discretization step* $h = \delta x$ that implicitly occurs in the discrete operator $G(k)$.

More generally, let L be a partial differential operator depending on time and space. We shall take the following definition for a Cauchy problem associated with this operator:

Definition 2.1 Let φ be the solution of a Cauchy problem associated with operator $L\varphi = 0$) and let Ψ^k ($k = \delta t$) be that of the associated discrete problem $L^k\Psi^k = 0$ (with the above notation, and assuming that both these problems have a unique solution). We shall call *consistency error* the quantity

$$\mathcal{E}^k = L^k\varphi^k, \tag{43}$$

where φ^k is the projection on the mesh of the exact solution φ of problem (34)-(35). This consistency error is a vector $\mathcal{E}^k = \left(\varepsilon^n\right)_{n\geq 1}$, each of whose components depends on $h = \delta x$ and on $k = \delta t$. The scheme is called *consistent* if for all n, $\varepsilon^n \to 0$, as $k \to 0$ and $h \to 0$, for a given choice of norm $||.||$ in space. If moreover we have,

$$\forall n, \quad ||\varepsilon^n|| = 0(k^p) + O(h^q), \tag{44}$$

the scheme is said to be of *order p in time and q in space* (for this norm).

We note that, by virtue of (41) and (37), we have:

$$\varepsilon^n = \frac{S(k) - G(k)}{k}\,\varphi(t^{n-1}). \tag{45}$$

The notion of *consistency* enables us to measure the error produced by *approximating the continuous operator* by a discrete operator. It can be computed on the *exact solution* of the continuous problem, thanks to a Taylor expansion. However, this will not be enough to let us prove the convergence of the scheme. Another notion is required, that of *stability*, which we shall now define.

3.3 Stability of scheme

By an immediate induction from (38)-(39), we obtain

$$\psi^n = G(k)^n \psi^0, \tag{46}$$

i.e. the approximate solution at time t^n is defined as a function of the initial condition through the operator $G(k)^n$.

Stability will force these operators to remain bounded when $k = \delta t \to 0$, $n \to \infty$, with the product $n\delta t$ remaining bounded by the final time T. The idea is that *there can be no growth over time*: the approximate solution must remain bounded, despite the accumulation of discretization and roundoff errors.

Definition 2.2 The scheme (38)-(39) is said to be *stable* if $G(k)^n$ remains *uniformly bounded* for all $k = \delta t$, n satisfying:

$$0 \leq k \leq k^*, \quad 0 \leq nk \leq T; \tag{47}$$

or, in other words, if there exists a positive constant $C_{st}(T)$ such that

$$\forall n, \forall k \in]0, k^*], \; 0 \leq nk \leq T, \qquad \text{we have: } ||[G(k)]^n|| \leq C_{st}(T), \tag{48}$$

for a given choice of norm in space.

We shall come back later, on concrete examples, to the practical way of checking stability. We shall now see how this notion is an essential features in the convergence of the scheme.

3.4 Convergence of scheme

The question we now ask is the following: in what sense will the approximate solution $\Psi^k = (\psi^n)_{n\geq 0}$ ($k = \delta t$) defined by (38)-(39) "tend" towards the exact solution of the initial problem (34)-(35) when the space step $h = \delta x$ and the time step $k = \delta t$ both go to 0? The answer lies in the following theorem, usually called the "Lax Equivalence Theorem".

Theorem 2.3 *If scheme (38)-(39) is stable and consistent, then it is convergent, that is the the error $e^n = \varphi(t^n) - \psi^n$ at time t^n goes to zero when the time and space steps both go to 0 (for the norm used in definitions 2.1 and 2.2), with the constraint $0 \leq n\delta t \leq T$.*

Proof Let us denote by E^k the error "vector" whose index n entry is the error e^n. By the definition of ψ^0, we have $e^0 = 0$. Furthermore, since $L^k \Psi^k = 0$, we obtain, with notation as in (43),

$$L^k E^k = L^k \varphi^k = \mathcal{E}^k, \tag{49}$$

and this means, according to (41), that

$$\left(L^k E^k \right)^n = \frac{e^n - G(k) e^{n-1}}{k} = \varepsilon^n, \tag{50}$$

or that:

$$e^n = G(k)e^{n-1} + k\varepsilon^n. \tag{51}$$

By an immediate induction, it follows that:

$$e^n = [G(k)]^n e^0 + k \sum_{i=1}^{n} [G(k)]^{n-i} \varepsilon^i; \tag{52}$$

then, as the error at the initial time e^0 is zero, we obtain the following estimate, for all integers n and all time steps $k = \delta t$ satisfying the constraint $0 \leq nk \leq T$,

$$||e^n|| \leq nk\, C_{st} \max_{i \in \{1,\ldots,n\}} |\varepsilon^i|, \tag{53}$$

because of the stability condition (48). We deduce that

$$||e^n|| \leq T C_{st} \max_{i \in \{1,\ldots,n\}} |\varepsilon^i|, \tag{54}$$

and because the scheme is consistent, this goes to 0 with the space and time steps.

Remark 2.4 All the above estimates are for a given choice of norm in space, and for a discretization scheme (finite differences, finite elements,...) yet to be determined.

Remark 2.5 We could also consider the case of an equation of the type $L\varphi = f$, with a source term f depending only on time; this would not change the above stability analysis, since this additional term would only affect the consistency error. The convergence theorem remains valid in that case.

Now that we have defined all these notions, we shall apply them to the study of some classical examples for the approximation of certain "model" evolution problems, starting with the example of the heat equation.

4 The heat equation

We present and analyze here several schemes to approximate the heat equation, using finite differences in both time and space. We start our study with the case of only one space variable in the whole space R^d:

$$\frac{\partial \varphi}{\partial t} - \frac{\partial^2 \varphi}{\partial x^2} = 0, \quad x \in R, \quad 0 < t \leq T, \tag{55}$$

$$\varphi(x, 0) = \varphi^0(x). \tag{56}$$

This problem is mathematically well-posed, and has a number of properties (Brezis, 1987), among them the *maximum principle* that we state without proof:

$$\text{if } \varphi^0 \geq 0, \quad \text{then} \quad 0 \leq \varphi(., t) \leq \sup_{x \in R} \varphi^0. \tag{57}$$

We shall denote by $\psi_i^n \simeq \varphi(x_i, t^n)$ the approximate solution taken at time $t^n = n\delta t$ ($n \in \{0, \ldots, M\}$, $M\delta t = T$) and at point $x_i = i\delta x$ ($i \in Z$). To simplify the notation we shall frequently write (k, h) for $(\delta t, \delta x)$.

4.1 An explicit scheme

We approximate the continuous operator

$$L = \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} \tag{58}$$

by the operator discretized *in space and time*, denoted by $\bar{L}$ for simplicity's sake, defined on sequences $\bar{\psi} = (\psi_i^n)_{i\in Z, n\in\{0,\dots,M\}}$ by:

$$\begin{aligned} &\bar{L}\bar{\psi} = \left((\bar{L}\bar{\psi})_i^n \right)_{i\in Z^+, n\in\{1,\dots,M\}}, \forall i \in Z,\ \forall n \in \{0,\dots,M-1\}, \\ &(\bar{L}\bar{\psi})_i^{n+1} = \frac{\psi_i^{n+1} - \psi_i^n}{k} - \frac{\psi_{i+1}^n - 2\psi_i^n + \psi_{i-1}^n}{h^2}. \end{aligned} \tag{59}$$

Then the discrete problem is written as:

$$\bar{L}\bar{\psi} = 0, \tag{60}$$

$$\forall i \in Z, \quad \bar{\psi}_i^0 = \varphi^0(x_i). \tag{61}$$

This scheme is *fully explicit*, *i.e.* given the approximate solution at step n (ψ_i^n known, $\forall i \in Z$), we obtain the approximate solution at step $n+1$ from the very simple relation

$$\forall i \in Z, \quad \psi_i^{n+1} = \psi_i^n + \frac{k}{h^2}(\psi_{i+1}^n - 2\psi_i^n + \psi_{i-1}^n), \tag{62}$$

which shows, in particular, that problem (60)-(61) admits a unique solution. This scheme is called the *forward Euler scheme*.

4.1.1 Convergence in the l^∞ norm in space

As far as the *consistency* error is concerned, we have the following result:

Proposition 3.1 *If the solution of the continuous problem (55)-(56) is C^2 in time and C^4 in space, then the scheme (59)-(61) is consistent and of order 1 in time and 2 is space (for the norm $||.||_\infty$ in space).*

Proof The consistency error is defined as $\bar{\varepsilon} = \bar{L}\bar{\varphi} = \bar{L}\bar{\varphi} - L\varphi$, where $\bar{\varphi}$ is the projection on the mesh of the exact solution φ at each node of the mesh in both time and space. It is a doubly indexed sequence $\bar{\varepsilon} = (\varepsilon_i^n)_{i\in Z^+, n\in\{1,\dots,M\}}$ defined by:

$$\varepsilon_i^{n+1} = [\tfrac{\varphi(x_i,t^{n+1}) - \varphi(x_i,t^n)}{k} - \tfrac{\partial\varphi}{\partial t}(x_i,t^n)] \tag{63}$$

$$-[\tfrac{\varphi(x_{i+1},t^n) - 2\varphi(x_i,t^n) + \varphi(x_{i-1},t^n)}{h^2} - \tfrac{\partial^2\varphi}{\partial x^2}(x_i,t^n)]. \tag{64}$$

The consistency error is thus the sum of two errors, the first one linked to the discretization of the time derivative operator, and the second one, relative to the spatial derivative, that has already been estimated in proposition 1.1. A simple Taylor expansion up to second order now gives, because of ({ref1.10):

$$\varepsilon_i^{n+1} = \frac{k}{2}\frac{\partial^2\varphi}{\partial t^2}(x_i,\tau^n) - \frac{h^2}{12}\frac{\partial^4\varphi}{\partial x^4}(\xi_i,t^n), \tag{65}$$

with $\tau^n \in]t^n, t^{n+1}[$ and $\xi_i \in]x_{i-1}, x_{i+1}[$. Thus, modulo the smoothness hypotheses in the statement of the theorem, we deduce that there exists two positive constants C_1 and C_2 such that, for all indices $i \in Z$ et $n \in \{0, ..., M-1\}$, we have

$$|\varepsilon_i^{n+1}| \leq C_1 k + C_2 h^2, \tag{66}$$

which shows that the scheme is consistent, of first order in time and second order in space; the above constant are defined by:

$$C_1 = \sup_{(x,t)\in R\times[0,T]} \frac{\partial^2 \varphi}{\partial t^2}(x,t), \quad C_2 = \sup_{(x,t)\in R\times[0,T]} \frac{\partial^4 \varphi}{\partial x^4}(x,t). \tag{67}$$

Let us now study the stability of the scheme in the l^∞ norm in space. Let us set:

$$\lambda = \frac{k}{h^2} \geq 0 \ ; \tag{68}$$

equality (62) then becomes

$$\forall i \in Z, \quad \psi_i^{n+1} = \lambda \psi_{i+1}^n + (1-2\lambda)\psi_i^n + \lambda \psi_{i-1}^n, \tag{69}$$

i.e. ψ_i^{n+1} is a linear combination of ψ_{i+1}^n, ψ_i^n and ψ_{i-1}^n. We then note that, if the following condition is satisfied

$$0 \leq \lambda = \frac{k}{h^2} \leq \frac{1}{2}, \tag{70}$$

all the coefficients of this linear combination are positive, and their sum is 1, so that

$$\forall i \in Z, \quad |\psi_i^{n+1}| \leq ||\Psi^n||_\infty, \tag{71}$$

if we denote by Ψ^n the vector in R^{Z^+} whose entries are ψ_i^n. By an immediate induction, we obtain

$$\forall n \geq 0, \quad ||\Psi^n||_\infty \leq ||\Psi^0||_\infty, \tag{72}$$

and this proves the l^∞ stability of the scheme.

We have just proved the following result:

Proposition 3.2 *Under condition (70), scheme (59)-(61) is stable for the norm $||.||_\infty$ in space.*

Propositions 3.1 and 3.2 completed by theorem 2.3 allow us to conclude that scheme (59)-(61) is convergent under condition (70).

Remark 3.3 By an immediate induction, we note that, if Ψ^0 is positive (meaning that all its components are positive), then this is also true for vector Ψ^n, still assuming hypothesis (70). We find here a discrete version of the maximum principle (57) seen at the beginning of this Section.

Remark 3.4 The stability condition (70) we found is seen here as a *sufficient* condition for stability. Is is also necessary? To answer this question, we shall study stability in a different context. This is what we do in the next Section.

4.1.2 Von Neumann stability

To make the practical study of stability simpler, we shall change (59)-(60) into a "continuous" in space version written as:

$$\forall x \in R, \quad \frac{\psi^{n+1}(x) - \psi^n(x)}{k} - \frac{\psi^n(x+h) - 2\psi^n(x) + \psi^n(x-h)}{h^2} = 0. \tag{73}$$

Let us denote by $\hat{\psi}$ the *Fourier transform* of ψ defined by:

$$\hat{\psi}(\xi) = \int_{-\infty}^{+\infty} e^{-i\xi x}\psi(x)dx. \tag{74}$$

Let us recall that:

$$\hat{\hat{\psi}} = 2\pi\psi, \quad ||\hat{\psi}||_2 = \sqrt{2\pi}\,||\psi||_2, \tag{75}$$

if we denote by $||.||_2$ the norm in the space $L^2(R)$; the second property is called "Plancherel's theorem". Let us Fourier transform equation (73); after using a change of variables to observe that

$$\int_{-\infty}^{+\infty} e^{-i\xi x}\varphi(x+h)dx = e^{i\xi h}\int_{-\infty}^{+\infty} e^{-i\xi y}\varphi(y)dy, \tag{76}$$

we obtain

$$\hat{\psi}^{n+1}(\xi) = \hat{\psi}^n(\xi) + \frac{k}{h^2}(e^{ih\xi}\hat{\psi}^n(\xi) - 2\hat{\psi}^n(\xi) + e^{-ih\xi}\hat{\psi}^n(\xi)). \tag{77}$$

This relation can be written

$$\hat{\psi}^{n+1}(\xi) = a(\xi)\hat{\psi}^n(\xi), \tag{78}$$

where the factor $a(\xi)$, called the *amplification factor*, is the real number defined by:

$$a(\xi) = 1 + \frac{k}{h^2}(e^{ih\xi} - 2 + e^{-ih\xi}) = 1 - 4\frac{k}{h^2}\sin^2\frac{h\xi}{2}. \tag{79}$$

By an immediate induction, we obtain

$$\hat{\psi}^n(\xi) = [a(\xi)]^n\hat{\psi}^0(\xi) = [a(\xi)]^n\hat{\varphi}^0(\xi). \tag{80}$$

If the following condition is satisfied

$$||a||_\infty = \sup_{\xi\in R}|a(\xi)| \leq 1, \tag{81}$$

we have, by using Plancherel's relation

$$||\psi^n||_2 = \frac{1}{\sqrt{2\pi}}||\hat{\psi}^n||_2 \leq \frac{1}{\sqrt{2\pi}}||\hat{\varphi}^0||_2 = ||\varphi^0||_2, \tag{82}$$

which proves the stability of the scheme for the norm $||.||_2$. Condition (81) thus appears as a sufficient condition for the stability of scheme (73) in the norm $||.||_2$. Is the condition necessary?

Before we answer this question, let us try and link this condition with the one we found previously. We remark that $a(\xi)$ is always less than 1, so that (81) is equivalent to $\forall\xi \in R$, $a(\xi) \geq -1$, and this relation is equivalent to (70).

graph of function $A(t) = a(\xi)$, $t = h\xi$

On figure4.1.2, we have plotted the graph of the function $A(t) = a(\xi)$, $t = h\xi$, for three different values of the ratio k/h^2: the curve with diamond-shaped symbols corresponds to $k/h^2 = 1/4$, that with crosses corresponds to the limit case $k/h^2 = 1/2$, and last the solid line curve corresponds to $k/h^2 = 1$. As the theoretical study predicts, condition (81) is not satisfied in the last case, whereas it is satisfied in the other cases.

We shall now prove the following result, in a slightly more general framework than is needed for the scheme (73):

Proposition 3.5 *A scheme of the type $\psi^n = G(k)^n \varphi^0$ ($k = \delta t$), where the Fourier transform of the operator $G(k)$ is a multiplication operator by a scalar function a, is stable in the $L^2(R)$ norm (we also say "von Neumann stable") if and only if the following stability condition, called "von Neumann condition", is satisfied:*

$$\exists C \geq 0, \exists (\delta t)^*, \quad \text{such that: } \forall \delta t \in]0, (\delta t)^*], \forall \xi \in R, |a(\xi)| \leq 1 + C\delta t. \tag{83}$$

To prove this result, we shall use a lemma that we present here in the general vector case (*i.e.* in R^d with $d \geq 1$), because we will find it useful later. Before we do that, we shall recall some definitions and results from matrix algebra [Ciarlet (1988), Lascaux-Théodor (1987), Schatzman (1991)].

Let B be a matrix with real or complex entries, of size $d \times d$, and let $B^* = \bar{B}^T$ be its transconjugate matrix; we say B is a *normal* matrix if it commutes with B^*, *i.e.* if $BB^* = B^*B$. Real symmetric matrices are normal. The *spectral radius* of matrix B, denoted by $\rho(B)$, is the largest modulus of all the eigenvalues of B. For any vector norm, it is possible to define a matrix norm, called the *subordinate matrix norm* to this vector norm, through the relation:

$$||B|| = \sup_{x \in R^d, x \neq 0} \frac{||Bx||}{||x||}; \tag{84}$$

for such a matrix norm, we have:

$$\forall B, \quad \rho(B) \leq ||B||. \tag{85}$$

In particular, the matrix norm subordinate to the Euclidean norm $||.||_2$ is defined by:

$$||B||_2 = \sqrt{\rho(BB^*)}. \tag{86}$$

If the matrix B is normal, its Euclidean norm is equal to its spectral radius.

After these reminders, we now state the following classical result, a proof of which can be found, for example, in Lascaux (1976), Richtmyer-Morton (1967).

Lemma 3.6 *Let F be the operator defined over $L^2(R^d)$ by: $\forall U \in R^d$, $F(U) = AU$, where, for all $\xi \in R^d$, $A(\xi)$ is a matrix; this linear operator*

has norm (we write L^2_d for $L^2(R^d)$)

$$||F||_{\mathcal{L}(L^2_d,L^2_d)} = \sup_{\xi\in R} ||A(\xi)||_2. \tag{87}$$

Proof of Proposition 3.5 Let us consider a scheme of the form $\psi^n = G(k)^n\varphi^0$, written, after Fourier transform, as $\hat{\psi^n} = F(\hat{\varphi^0})$, where the operator F is the multiplication operator by the scalar function a^n. If we use first Plancherel's Theorem and then Lemma 3.6 in the scalar case (*i.e.* $d = 1$), we deduce

$$|||[G(k)]^n||_{\mathcal{L}(L^2(R_x),L^2(R_x))} = ||F||_{\mathcal{L}(L^2(R_\xi),L^2(R_\xi))} = ||a^n||_\infty = ||a||^n_\infty, \tag{88}$$

with the notation: $||a||_\infty = \sup_{\xi\in R} |a(\xi)|$.

Let us assume that the scheme is stable in $L^2(R)$. By definition, the norms of the operators $G(k)^n$, as operators from $L^2(R)$ into $L^2(R)$, remain bounded independently of n, for all integers n and for time steps $k = \delta t \in]0, k^*]$ satisfying the constraint $nk \leq T$. There exists a positive constant C_1 (we may assume $C_1 > 1$) such that $||a||^n_\infty \leq C_1$. This holds in particular for the integer $n = n_0$, where n_0 is half the integer part of the ration $T/k \geq 1$. Thus, we have

$$||a||_\infty \leq C_1^{2\frac{k}{T}} \leq 1 + \frac{C_1^{2\frac{k^*}{T}} - 1}{k^*}\, k, \tag{89}$$

or (83) with $C = (C_1^{2k^*/T} - 1)/k^*$.

Conversely, if a satisfies condition (83), then

$$||a||^n_\infty \leq (1 + Ck)^n \leq (e^{Ck})^n \leq e^{CT}, \tag{90}$$

for all integers n and all time steps $k = \delta t$ such that $nk \leq T$, which proves the L^2 stability of the scheme and ends the proof.

Remark 3.7 The term $C\delta t$ in (83) allows an exponential growth in time of the numerical solution (while noting that this growth is actually limited by the fact that the definition of stability is only concerned with solutions in finite time T).

In general, this condition is often replaced by the more restrictive condition (81), called "strict von Neumann condition"; this is true, in particular, if we know that the exact solution has no exponential behavior in time, as is precisely the case here. Under this condition, the scheme in proposition (59) is "strictly stable": the numerical solution does not grow faster than the exact solution, and we have stability in the limit case $T = +\infty$. From a practical point of view, it is also preferable to use condition (81); indeed, even though the theoretical condition (83) is true in the limit $\delta t \to 0$, because the discretization steps have a nonzero finite value, the numerical solution may grow appreciably during the time iterations, harming the effective stability of the scheme.

However, there are situations (Richtmyer-Morton (1967)) where the condition (83) is theoretically indispensable; this would for instance be the case if

equation (55) featured a dissipation term like $b\varphi$ discretized explicitly (*i.e.* by $b\psi_i^n$ at point x_i).

One shows easily that the scheme (73) is always consistent, of order 1 in time and 2 in space, and this eventually allows us to state the final convergence result, whose proof is analogous to that of theorem 2.3.

Theorem 3.8 *Under the stability condition (70), the scheme (73) is convergent for the $L^2(R)$ norm; it is of order 1 in time and 2 in space.*

Remark 3.9 This Fourier transform method is an undeniably practical tool to study the stability of schemes, that also gives necessary and sufficient stability conditions: we shall use it as often as possible. However, its major drawback is that it is difficult to generalize to the case of a boundary value problem, and that it is limited to partial differential equations with constant coefficients discretized on a uniform mesh.

Remark 3.10 The stability condition (70) imposes a particularly severe constraint on the time step, since it must be of the order of the square of the space step (and even less!), and this means that in order to reach a given final time T, we shall have to iterate the algorithm a large number of times, which leads to prohibitive computer times. This fully explicit scheme, even though it is very simple to program, is thus not a very good scheme. We shall now propose others, and study their properties somewhat more rapidly.

4.2 Towards implicit schemes

When using scheme (59)-(61) to determine the approximate solution at step $n+1$ from that at step n, the Laplacian was computed at time step n. This had the advantage of making the programming at each time step quite simple, but made it very costly because of condition (70) on the time step. Conversely, what happens if the Laplacian is taken at time step $n+1$?

4.2.1 Study of the fully implicit scheme

The operator L defined by (58) is now approximated by the operator $\bar{L}$, discrete in time and space, that is defined on sequences $\bar{\psi} = (\psi_i^n)_{i\in Z^+, n\in\{0,\dots,M-1\}}$ by

$$\bar{L}\bar{\psi} = \left((\bar{L}\bar{\psi})_i^n \right)_{i\in Z^+, n\in\{1,\dots,M\}}, \tag{91}$$

$$\forall i \in Z,\ \forall n \in \{0,\dots,M-1\}, \tag{92}$$

$$(\bar{L}\bar{\psi})_i^{n+1} = \frac{\psi_i^{n+1} - \psi_i^n}{k} - \frac{\psi_{i+1}^{n+1} - 2\psi_i^{n+1} + \psi_{i-1}^{n+1}}{h^2}, \tag{93}$$

where h still denotes the space step and k is the time step. The resulting scheme $\bar{L}\bar{\psi} = 0$ is said to be *implicit* because, when we know the approximate solution at time t^n, the solution at the next time step (given by relation (91)) is not determined in a simple way, since it requires *solving a non diagonal linear system*. We shall come back to this point in more details when we treat the case

of the heat equation in a bounded domain in R. Before we do that, we shall study the stability properties of such a scheme.

Analogously to (73), the "continuous in space" version of this scheme is written as:

$$\forall x \in R, \quad \frac{\psi^{n+1}(x) - \psi^n(x)}{k} - \frac{\psi^{n+1}(x+h) - 2\psi^{n+1}(x) + \psi^{n+1}(x-h)}{h^2} = 0, \tag{94}$$

which enables us to study its von Neumann stability. The Fourier transform $\hat{\psi}$ of ψ is now a solution of

$$\hat{\psi}^{n+1}(\xi) = \hat{\psi}^n(\xi) + \frac{k}{h^2}(\hat{\psi}^{n+1}(\xi)e^{ih\xi} - 2\hat{\psi}^{n+1}(\xi) + e^{-ih\xi}\hat{\psi}^{n+1}(\xi)) \; ; \tag{95}$$

in other words

$$\hat{\psi}^{n+1}(\xi) = b(\xi)\hat{\psi}^n(\xi), \tag{96}$$

where the *amplification factor* $b(\xi)$ is now defined by:

$$b(\xi) = \frac{1}{1 + 4\frac{k}{h^2}\sin^2\frac{h\xi}{2}}. \tag{97}$$

We notice that

$$\forall h \geq 0, \forall k \geq 0, \quad 0 \leq b(\xi) \leq 1, \tag{98}$$

which allows us to obtain by induction and Plancherel's theorem the following inequalities

$$||\psi^n||_{L^2} \leq ||\psi^{n-1}||_{L^2} \leq \ldots \leq ||\psi^1||_{L^2} \leq ||\varphi^0||_{L^2}, \tag{99}$$

showing that the scheme is stable, without any restrictive condition on the time step; we say the scheme is *unconditionally stable*. We have just proven

Proposition 3.11 *The fully implicit scheme is unconditionally von Neumann stable.*

On figure VII.3, we have plotted the graph of the function $B(t) = b(\xi)$, $t = h\xi$, for the three values of the ration k/h^2 already considered in figure VII.2: the curve with the diamond shaped symbols corresponds to $k/h^2 = 1/4$, that with crosses is the limit case $k/h^2 = 1/2$, and the solid line curve corresponds to $k/h^2 = 1$. Relation (98) is satisfied in all three cases.

Remark By doing a Taylor expansion around point (x, t^{n+1}), one can show that the consistency error is the same as that of the explicit scheme.

This scheme, called the *backward Euler scheme* is very robust, but it is still only of first order in time. We shall now try and construct other stable schemes, hopefully more accurate, by using a linear combination with the explicit scheme.

graph of function $B(t) = b(\xi)$, $t = h\xi$

4.2.2 Theta schemes

Let θ be a fixed parameter in $[0,1]$; we define the θ-scheme by

$$\frac{\psi_i^{n+1}-\psi_i^n}{k} - \mathcal{D}_i^+\mathcal{D}_i^-\left(\theta\Psi^{n+1}+(1-\theta)\Psi^n\right)\big) = 0, \tag{100}$$

where Ψ^n denotes the vector in R^{Z^+} with entries ψ_i^n, and where $\mathcal{D}^+$ and $\mathcal{D}^-$ are non-centered operators in space defined on sequences $U=(u_i)_{i\in Z^+}$ of R^{Z^+} by a relation analogous to definitions (6) and (7) for functions, that is:

$$\mathcal{D}_i^+U = \frac{1}{h}(u_{i+1}-u_i), \quad \mathcal{D}_i^-U = \frac{1}{h}(u_i-u_{i-1}); \tag{101}$$

by analogy with (11), we have obviously:

$$\mathcal{D}_i^+\mathcal{D}_i^-U = \mathcal{D}_i^-\mathcal{D}_i^+U = \frac{u_{i+1}-2u_i+u_{i-1}}{h^2}. \tag{102}$$

For $\theta=0$, this scheme is the explicit scheme, whereas for $\theta=1$, it is the implicit scheme. We shall see that $\theta=1/2$ plays a particular role: in that case the scheme is called *the Crank-Nicolson scheme.*

By Fourier transform in space, the scheme $\forall x\in R$,

$$\frac{\psi^{n+1}(x)-\psi^n(x)}{k} - \theta\,\frac{\psi^{n+1}(x+h)-2\psi^{n+1}(x)+\psi^{n+1}(x-h)}{h^2} \tag{103}$$

$$-(1-\theta)\,\frac{\psi^n(x+h)-2\psi^n(x)+\psi^n(x-h)}{h^2} = 0, \tag{104}$$

becomes

$$\hat{\psi}^{n+1}(\xi) = c(\xi)\hat{\psi}^n(\xi), \tag{105}$$

and the *amplification factor* $c(\xi)$ is now equal to:

$$c(\xi) = \frac{1-4(1-\theta)\frac{k}{h^2}\sin^2\frac{h\xi}{2}}{1+4\theta\frac{k}{h^2}\sin^2\frac{h\xi}{2}}. \tag{106}$$

graph of $C(t)=c(\xi)$, $t=h\xi$, for $\theta=1/2$

We notice that for all real values of ξ we have $c(\xi)\le 1$, so that the scheme is von Neumann stable if and only if: $\forall\xi\in R$, $c(\xi)\ge -1$. This last property may be written

$$\forall\xi\in R, \quad 2-4(1-2\theta)\frac{k}{h^2}\sin^2\frac{h\xi}{2}\ge 0, \tag{107}$$

and this relation is always satisfied for $\theta\ge 1/2$ (but this is not a necessary and sufficient condition).

graph of $C(t)=c(\xi)$, $t=h\xi$, for $\theta=1/4$

We thus have

Proposition 3.12 *For $\theta \geq 1/2$, the θ-scheme is unconditionally von Neumann stable.*

graph of $C(t) = c(\xi)$, $t = h\xi$, for $\theta = 1/8$

On figures 4.2.2,4.2.2,4.2.2, we have plotted the graph of the function $C(t) = c(\xi)$, $t = h\xi$, for the same three values of the ratio k/h^2 considered in figures VII.2 and VII.3 above (the symbol captions are the same as on those figures), for three different values of the parameter θ: first for $\theta = 1/2$, then for $\theta = 1/4$, and last for $\theta = 1/8$. Figures VII.4 and VII.6 confirm the theoretical results; for $\theta = 1/4$, the scheme is stable if $k/h^2 \leq 1$, and this is precisely the case on figure VII.5.

Let us now show a *consistency* property peculiar to the Crank-Nicolson scheme, as all the others have the same accuracy characteristics as the above schemes.

Proposition 3.13 *If the solution of the continuous problem (55)-(56) is sufficiently smooth (C^3 in time and C^4 in space), the Crank-Nicolson scheme ($\theta = 1/2$) is consistent of order 2 in time and space.*

Proof The proof follows from a Taylor expansion of the exact solution φ of (55)-(56) at point $x_i = ih$ in space and $t^n + k/2$ ($t^n = nk$) in time.

For simplicity's sake, let us denote by $\tilde{\varphi}(x_i, .)$ the function of the sole time variable defined by the following difference quotient:

$$\tilde{\varphi}(x_i, t) = \frac{\varphi(x_{i+1}, t) - 2\varphi(x_i, t) + \varphi(x_{i-1}, t)}{h^2}. \tag{108}$$

According to proposition 1.1, we already know that if φ is of class C^4 in space, we have for any time t:

$$\tilde{\varphi}(x_i, t) = \frac{\partial^2 \varphi}{\partial x^2}(x_i, t) + 0(h^2). \tag{109}$$

The consistency error $\bar{\varepsilon} = \bar{L}\bar{\varphi}$, where $\bar{\varphi}$ is the projection of the exact solution φ at each of the nodes of the mesh in time and space, is a doubly indexed sequence $\bar{\varepsilon} = (\varepsilon_i^n)_{i \in Z^+, n \in \{1,\ldots,M\}}$ that cab be defined by

$$\varepsilon_i^{n+1} = [\tfrac{\varphi(x_i, t^{n+1}) - \varphi(x_i, t^n)}{k} - \tfrac{\partial \varphi}{\partial t}(x_i, t^{n+1/2})] \tag{110}$$

$$-[\tfrac{1}{2}\,(\,\tilde{\varphi}(x_i, t^{n+1}) + \tilde{\varphi}(x_i, t^n)\,) - \tfrac{\partial^2 \varphi}{\partial x^2}(x_i, t^{n+1/2})], \tag{111}$$

since $(L\varphi)(., t^{n+1/2}) = 0$. The consistency error is thus the sum of two errors $\varepsilon_i^{n+1} = (\varepsilon_t)_i^{n+1} + (\varepsilon_x)_i^{n+1}$, one

$$(\varepsilon_t)_i^{n+1} = \frac{\varphi(x_i, t^{n+1}) - \varphi(x_i, t^n)}{k} - \frac{\partial \varphi}{\partial t}(x_i, t^{n+1/2}), \tag{112}$$

linked to the discretization of the time derivative operator, whereas the other

$$(\varepsilon_x)_i^{n+1} = \frac{1}{2}\,[\,\tilde{\varphi}(x_i, t^{n+1}) + \tilde{\varphi}(x_i, t^n)\,] - \frac{\partial^2 \varphi}{\partial x^2}(x_i, t^{n+1/2}), \tag{113}$$

is linked to the discretization of the spatial operator.

Let now u be any function of the t variable, assumed of class C^3; we have the Taylor expansions:

$$u(t+k) = u(t+\frac{k}{2}) + \frac{k}{2}\frac{\partial u}{\partial t}(t+\frac{k}{2}) + \frac{k^2}{8}\frac{\partial^2 u}{\partial t^2}(t+\frac{k}{2}) + O(k^3), \tag{114}$$

$$u(t) = u(t+\frac{k}{2}) - \frac{k}{2}\frac{\partial u}{\partial t}(t+\frac{k}{2}) + \frac{k^2}{8}\frac{\partial^2 u}{\partial t^2}(t+\frac{k}{2}) + O(k^3). \tag{115}$$

If we apply these expansions to $u = \varphi(x_i, .)$, we obtain by subtracting that:

$$(\varepsilon_t)_i^{n+1} = O(k^2). \tag{116}$$

According to (109), the error ε_x can be written as

$$(\varepsilon_x)_i^{n+1} = \frac{1}{2}\,[\,\frac{\partial^2 \varphi}{\partial x^2}(x_i, t^{n+1}) + \frac{\partial^2 \varphi}{\partial x^2}(x_i, t^n)\,] - \frac{\partial^2 \varphi}{\partial x^2}(x_i, t^{n+1/2}) + O(h^2). \tag{117}$$

Then it suffices to apply (114) to the function $u = \frac{\partial^2 \varphi}{\partial x^2}(x_i, .)$ to deduce that

$$(\varepsilon_t)_i^{n+1} = O(k^2) + O(h^2), \tag{118}$$

and this ends the proof.

Remark 3.14 If equation (55) has a source term f that only depends on time, this term only occurs in the consistency error, and not in stability (*cf.* remark 2.5). In order to keep second order accuracy in time, we must consider the scheme

$$\frac{\psi_i^{n+1} - \psi_i^n}{k} - \mathcal{D}_i^+ \mathcal{D}_i^- (\theta \Psi^{n+1} + (1-\theta)\Psi^n)) = \frac{f(t^{n+1}) + f(t^n)}{2}, \tag{119}$$

because

$$\frac{f(t^{n+1}) + f(t^n)}{2} = f(t^n + \frac{k}{2}) + O(k^2), \tag{120}$$

and this is again true by (114).

4.3 A three level scheme

Let us go back to the simple example of the explicit scheme. The question we ask is the following: couldn't we improve the discretization in time by taking a second order approximation of the operator $\frac{\partial}{\partial t}$?

A rather natural choice consists for example in considering the *Richardson scheme*:

$$\forall n \geq 1, \quad \frac{\psi_i^{n+1} - \psi_i^{n-1}}{2k} - \frac{\psi_{i+1}^n - 2\psi_i^n + \psi_{i-1}^n}{h^2} = 0. \tag{121}$$

As above, we require at the initial time: $\psi_i^0 = \varphi(x_i), \forall i$. It is clear that we cannot use (61) to define the approximate solution at the next step; we shall then use a first order scheme, of the explicit type, to compute Ψ^1.

Proposition 3.15 *Richardson's scheme is, modulo enough smoothness for the solution of (55)-(56), second order in time and space; unfortunately, it is always unstable.*

Proof We leave it as an exercise to the reader to compute the consistency error of this scheme. Let us study the *von Neumann stability*. Again using the Fourier transform, we have:

$$\hat{\psi}^{n+1}(\xi) = d(\xi)\hat{\psi}^n(\xi) + \hat{\psi}^{n-1}(\xi), \tag{122}$$

with

$$d(\xi) = -8\frac{k}{h^2}\sin^2\frac{h\xi}{2}. \tag{123}$$

Equation (122) can be written in matrix form

$$\hat{X}^{n+1}(\xi) = A(\xi)\hat{X}^n(\xi), \tag{124}$$

where we denote by X^n the vector with entries $(\psi^{n+1}, \psi^n)^T$, and A the *amplification matrix*:

$$A(\xi) = \begin{bmatrix} d(\xi) & 1 \\ 1 & 0 \end{bmatrix}. \tag{125}$$

Since A is symmetric its eigenvalues are real. They are solutions of the characteristic equation $\lambda^2 - \lambda d(\xi) - 1 = 0$. But the product of the roots is -1, so there must exists values of ξ for which one of the eigenvalues has absolute value strictly larger than 1 (they cannot both have absolute value 1, because for $\lambda = 1$ or -1, the above characteristic polynomial is not identically zero). We deduce that there are ξ values for which the spectral radius of $A(\xi)$ is strictly larger than 1, and since this matrix is normal, this contradicts the stability condition of the following Lemma (we cannot find a positive bounded constant C such that (126) holds): Richardson's scheme is thus always unstable.

For systems, we have the stability result summed up in the following way:

Lemma 3.16 *A necessary condition for a scheme whose Fourier transform is written in the form (124) to be stable is that there exists two positive constants C and k^* such that:*

$$\forall \xi \in R, \forall k \in]0, k^*[, \quad \rho(A(\xi)) \leq 1 + Ck. \tag{126}$$

If the matrix $A(\xi)$ is normal for all ξ, this von Neumann condition is a sufficient stability condition.

Proof We sketch the proof which is analogous to that of proposition 3.5. Using Lemma 3.6 in the case $d = 2$ and Plancherel's theorem, we show, thanks to the following relations

$$[\rho(A(\xi))]^n = \rho[(A(\xi))^n] \leq ||(A(\xi))^n||_2, \tag{127}$$

that (126) is a *necessary* for a scheme whose Fourier transform is written in the form (124) to be stable. Conversely, if the matrix $A(\xi)$ is normal, we have

$$||(A(\xi))^n||_2 \leq ||A(\xi)||_2^n = [\rho(A(\xi)]^n; \tag{128}$$

if moreover condition (126) is satisfied, we have, for all integers n such that $nk \leq T$,

$$||(A(\xi))^n||_2 \leq (1 + Ck)^n \leq e^{CT}, \tag{129}$$

which proves that the scheme is stable.

Remark 3.17 As we already said about the scalar case in Remark 3.7, the condition (126) on the spectral radius of the matrix A is usually replaced by the more restrictive condition

$$\forall \xi \in R, \quad \rho(A(\xi)) \leq 1, \tag{130}$$

called *strict von Neumann condition.*

One can find in Richtmyer-Morton (1967) other sufficient *stability* conditions; we state, without proof, the two main results, starting with the case of diagonalizable matrices. Let us recall that, given a matrix P whose columns are the components of a set of vectors, the Gram determinant of this set, denoted by Δ^2, is the determinant of the matrix P^*P; a necessary and sufficient condition for the vectors in this set to be linearly independent is that this Gram determinant be strictly positive.

Let us still consider schemes whose Fourier transform is written in the form (124) with a non necessarily normal matrix A; we have:

Theorem 3.18 *If the matrix A is diagonalizable and if there exists a constant $\delta > 0$ such that*

$$\forall \delta t \in]0, (\delta t)^*[, \forall \xi, \quad \Delta(\xi) \geq \delta > 0, \tag{131}$$

where $\Delta^2(\xi)$ is the Gram determinant of the normalized eigenvectors of the matrix $A(\xi)$, then the von Neumann condition is a necessary and sufficient stability condition.

Theorem 3.19 *If the elements of the matrix A are bounded for all $\delta t \in]0, (\delta t)^*[$ and for all ξ, and if all the eigenvalues $\lambda_i(\xi), i \in \{1, ..., N\}$ of $A(\xi)$, except possibly one of them, are inside the unit disk, i.e.*

$$\forall \delta t \in]0, (\delta t)^*[, \forall \xi, \quad |\lambda_1(\xi)| \leq 1, \tag{132}$$

$$\forall i \in \{2, ..., N\}, \forall \delta t \in]0, (\delta t)^*[, \forall \xi, \quad |\lambda_i(\xi)| \leq \gamma < 1, \tag{133}$$

then the scheme is stable.

4.4 Taking boundary conditions into account

4.4.1 Statement of the problem

In this Section we shall be concerned with discretizing the heat equation in a single space variable belonging to a bounded interval $\Omega =]0, L[$ in R, with Dirichlet boundary conditions:

$$\frac{\partial \varphi}{\partial t} - \frac{\partial^2 \varphi}{\partial x^2} = 0, \quad x \in \Omega =]0, L[, \quad 0 < t \leq T, \tag{134}$$

$$\varphi(x, 0) = \varphi^0(x), \tag{135}$$

$$\varphi(x, t) = g(x, t), \quad \text{for } x = 0 \text{ and } x = L \tag{136}$$

We shall assume that the initial condition φ^0 satisfies the boundary conditions (136) at $t = 0$ *i.e.* $\varphi^0(x) = g(x, 0)$ at the points $x = 0$ and $x = L$.

The difference with what we did previously lies in the way we treat the boundary conditions (136). The space step is now equal to $h = \delta x = L/(N+1)$, and that means that at any time t^n, there are N unknowns that are the values ψ_i^n of the approximate solution at the internal vertices $x_i = ih$, $i \in \{1, \ldots, N\}$ of the mesh; these values solve a "discretized" version, as described above, of equation (134). At both ends of the intervals, the solution is required to satisfy the boundary conditions (136), *i.e.*

$$\psi_i^n = g(x_i, t^n), \text{ for } i = 0 \text{ and } i = N + 1. \tag{137}$$

Two questions now arise:

- — how to analyze the consistency and the stability of the scheme, so as to ensure its convergence, as was the case in the whole space?
- — how to solve the discrete problem from a practical point of view?

We shall answer both questions successively for concrete schemes. As far as the first question is concerned, since the discrete solution satisfies the continuous boundary conditions at each node of the space-time mesh, the consistency error of the global scheme is that of the scheme used to discretized (134). There only remains to analyze its stability: this is what we shall now undertake.

The second question will be studied in section 3.5.

4.4.2 Stability by energy inequalities

Because we are in a bounded spatial domain, we can no longer use the Fourier transform to study stability. We shall first study the stability of the continuous problem, whence we shall deduce a method for the discrete problem. We assume, for simplicity's sake that the Dirichlet condition is homogeneous, *i.e.* $g = 0$.

At the continuous level, let us multiply equation (134) by φ, integrate the resulting equation over Ω, and integrate by parts; we obtain

$$\frac{1}{2} \frac{d}{dt} \Big(\int_\Omega \varphi^2(x, t) dx \Big) + \int_\Omega \Big(\frac{\partial \varphi}{dx} \Big)^2 (x, t) dx = 0, \tag{138}$$

which shows, since the second term is non-negative, that:

$$\frac{1}{2}\frac{d}{dt}\Big(\int_\Omega \varphi^2(x,t)dx\Big) \quad \leq 0. \tag{139}$$

In other words, the norm in $L^2(R)$ of $\varphi(.,t)$ decreases when time increases, because we have:

$$\forall t \geq 0, \quad ||\varphi(.,t)||_{L^2(R)} \leq ||\varphi^0||_{L^2(R)}. \tag{140}$$

This inequality, called *energy inequality*, shows the "stability in time" of the solution of the continuous problem (it also shows its uniqueness!). Let us note that inequality (140) is also valid for an homogeneous Neumann problem.

Remark 3.20 Let us point out that in the case of an non-homogeneous Dirichlet boundary condition, we obtain the same conclusion, after "lifting" (as we did for finite elements) the boundary condition, so as to reduce to a homogeneous problem.

We shall show on the example of the Crank-Nicolson scheme, how to copy closely this argument at the discrete level, so as to prove an energy inequality of the type

$$\forall n \geq 0,\ n\delta t \leq T, \quad ||\Psi^n||_{l^2} \leq ||\Psi^0||_{l^2}, \tag{141}$$

where we have set:

$$\Psi^n = (\psi_0^n, ..., \psi_{N+1}^n), \quad ||\Psi^n||_{l^2} = \sum_{i=0}^{N+1} (\psi_i^n)^2. \tag{142}$$

It is indeed clear, according to (46) and (48), that (141) proves the stability of the scheme for the l^2 norm in space.

The scheme we shall study is defined by (100) with $\theta = 1/2$ and (137), *i.e.*

$$\frac{\psi_i^{n+1} - \psi_i^n}{k} - \frac{1}{2}\mathcal{D}_i^+\mathcal{D}_i^-\left(\Psi^{n+1} + \Psi^n\right) = 0, \quad i \in \{1, ..., N\}, \tag{143}$$

$$\psi_i^n = 0, \text{ for } i = 0 \text{ and } i = N+1. \tag{144}$$

Proposition 3.21 *The Crank-Nicolson scheme (144)-(144) is unconditionally stable for the l^2 norm in space*

As in the continuous case, the proof rests on an "discrete integration by parts formula" that we shall prove first.

Lemma 3.22 *Let $U = (u_i)_{i\in\{0,...,N\}}$ and $V = (v_i)_{i\in\{0,...,N\}}$ be two sequences indexed by $\{0, ..., N\}$; then:*

$$\sum_{i=1}^{N} \mathcal{D}_i^+ U\, v_i \;=\; -\sum_{i=1}^{N+1} u_i\, \mathcal{D}_i^- V \;+ \frac{1}{h}\,(u_{N+1}v_{N+1} - u_1 v_0). \tag{145}$$

In particular, if $v_0 = v_{N+1} = 0$, *we have*:

$$\sum_{i=1}^{N} \mathcal{D}_i^+ U\, v_i \;=\; -\sum_{i=1}^{N+1} u_i\, \mathcal{D}_i^- V. \tag{146}$$

Proof Let us expand the left hand side of (34.54); we obtain successively

$$\sum_{i=1}^{N} \mathcal{D}_i^+ U v_i = \frac{1}{h}\sum_{i=1}^{N}(u_{i+1}-u_i)v_i = \frac{1}{h}(\sum_{i=1}^{N} u_{i+1}v_i - \sum_{i=1}^{N} u_i v_i) \tag{147}$$

$$= \frac{1}{h}(\sum_{i=2}^{N+1} u_i v_{i-1} - \sum_{i=1}^{N} u_i v_i) = \frac{1}{h}[\sum_{i=1}^{N+1} u_i(v_{i-1}-v_i) - u_1 v_0 + u_{N+1}v_{N+1}] \tag{148}$$

$$= -\sum_{i=1}^{N+1} u_i \mathcal{D}_i^- V \;+\; \frac{1}{h}(-u_1 v_0 + u_{N+1}v_{N+1}), \tag{149}$$

which proves (145) and ends the proof of the Lemma, (146) being a particular case of the above equality.

Proof of Proposition 3.21 Let us multiply equation (144) by $\psi_i^{n+1} + \psi_i^n$ and sum over all indices $i \in \{1, ..., N\}$; we obtain:

$$\sum_{i=1}^{N} \frac{(\psi_i^{n+1})^2 - (\psi_i^n)^2}{k} - \frac{1}{2}\sum_{i=1}^{N} \mathcal{D}_i^+\big(\mathcal{D}_i^-(\Psi^{n+1}+\Psi^n)\big)\,(\psi_i^{n+1}+\psi_i^n) = 0. \tag{150}$$

By virtue of the boundary conditions (144) satisfied by Ψ^n and Ψ^{n+1}, we can apply the relation (146) to the sequences $V = \Psi^{n+1} + \Psi^n$ and $U = (u_i)_i$, $u_i = \mathcal{D}_i^-(\Psi^{n+1} + \Psi^n)$, which gives us:

$$\sum_{i=1}^{N} \mathcal{D}_i^+\big(\mathcal{D}_i^-(\Psi^{n+1}+\Psi^n)\big)\,(\psi_i^{n+1}+\psi_i^n) \;=\; -\sum_{i=1}^{N+1}\big(\mathcal{D}_i^-(\Psi^{n+1}+\Psi^n)\big)^2 \;\le\; 0. \tag{151}$$

We thus deduce from (150) that

$$\sum_{i=1}^{N} \frac{(\psi_i^{n+1})^2 - (\psi_i^n)^2}{k} \;\le\; 0, \tag{152}$$

or

$$\sum_{i=1}^{N}(\psi_i^{n+1})^2 \;\le\; \sum_{i=1}^{N}(\psi_i^n)^2. \tag{153}$$

According to (144), we have, using notation as in (142):

$$\forall n \ge 0, \quad ||\Psi^{n+1}||_{l^2} \;\le ||\Psi^n||_{l^2}, \tag{154}$$

which gives, by an easy induction, a slightly stronger result than (141), since it allows the case $T = +\infty$; the l^2 stability of the scheme is thus proven.

We shall now show in detail, on an example, the practical solution of the scheme.

4.5 Practical solution of the implicit scheme

The discrete problem associated with problem (134)-(136) is particularly simple to solve if the scheme is explicit; we shall thus look at the implicit case, and to make the right hand side simpler, we shall consider the example of the fully implicit scheme defined by the discrete operator (91). The scheme we study can be written more generally ($f = 0$ in our example):

$$\frac{\psi_i^{n+1} - \psi_i^n}{k} - \frac{\psi_{i+1}^{n+1} - 2\psi_i^{n+1} + \psi_{i-1}^{n+1}}{h^2} = f_i^{n+1}, \quad i \in \{1, ..., N\}, \tag{155}$$

$$\psi_i^{n+1} = g(x_i, t^{n+1}), \text{ for } i = 0 \text{ and } i = N + 1, \tag{156}$$

which means that $X = (\psi_1^{n+1}, ..., \psi_N^{n+1})^T$ solves the linear system

$$AX = b, \tag{157}$$

where the matrix A is defined by

$$A = \begin{bmatrix} 2c+1 & -c & ... & 0 \\ -c & 2c+1 & -c & 0 \\ 0 & -c & 2c+1 & ... \\ & ... & ... & \\ ... & 0 & -c & 2c+1 \end{bmatrix}, \quad c = \frac{k}{h^2} > 0, \tag{158}$$

and the right hand side b is given by:

$$b = \begin{bmatrix} kf_1^{n+1} + \psi_1^n + cg(x_0, t^{n+1}) \\ kf_2^{n+1} + \psi_2^n \\ ... \\ kf_{N-1}^{n+1} + \psi_{N-1}^n \\ kf_N^{n+1} + \psi_N^n + cg(x_{N+1}, t^{n+1}) \end{bmatrix}. \tag{159}$$

We can easily show, as we did in Section 1.3, that the matrix A is symmetric and positive definite, which proves that the linear system (157), and thus the scheme (155)-(156) has one and only one solution.

We have at our disposal numerous methods enabling us to actually solve system (157); we shall give details for that based on the LU factorization of the matrix, which is particularly simple here, as A is tridiagonal.

We shall decompose A as a product of two matrices, a lower triangular one, L, and an upper triangular one, U, with unit diagonal; because of the sparse

structure of A, these two matrices L and U have only two nonzero diagonals. We thus have:

$$A = LU, \quad L = \begin{bmatrix} d_1 & & & 0 \\ l_1 & d_2 & & \\ & l_2 & d_3 & \\ & & \ldots & \\ 0 & & l_{N-1} & d_N \end{bmatrix}, \quad U = \begin{bmatrix} 1 & u_1 & & 0 \\ & 1 & u_2 & \\ & & \ldots & \\ & & 1 & u_{N-1} \\ 0 & & & 1 \end{bmatrix}. \tag{160}$$

The nonzero elements of these matrices are given by:

$$L_{i,i-1} = l_{i-1}, \quad L_{ii} = d_i, \quad U_{i,i+1} = u_i, \quad U_{ii} = 1. \tag{161}$$

With this notation, we have the following relations:

$$A_{ii} = 2c + 1 = (LU)_{ii} = L_{i,i-1}U_{i-1,i} + L_{ii}U_{ii} = u_{i-1}l_{i-1} + d_i, \tag{162}$$

$$A_{i,i-1} = -c = (LU)_{i,i-1} = L_{i,i-1}U_{i-1,i-1} = l_{i-1}, \tag{163}$$

$$A_{i,i+1} = -c = (LU)_{i,i+1} = L_{i,i}U_{i,i+1} = d_i u_i. \tag{164}$$

Once the matrices L and U have been computed, it is easy to solve system (157) in two consecutive steps, the first step, called "forward solve", where we solve the lower triangular system $Lz = b$, then the second step, called "backward solve", where we solve the upper triangular system $UX = z$. The algorithm is now the following, if we denote by X_i the entries of X and by b_i those of the right hand side b:

4.5.1 LU factorization algorithm

0 Set $d_1 = 2c + 1, \quad z_1 = b_1/d_1, \quad u_1 = -c/d_1$
1 Loop over $i = 2, \ldots, N$ (factorization and forward solve)
$l_{i-1} = -c$
$d_i = 2c + 1 - u_{i-1}l_{i-1}$
$u_i = -c/d_i$
$z_i = (b_i - l_{i-1}z_{i-1})/d_i$
2 Initialization of the backward solve by setting $X_N = z_N$
3 Loop over $j = N-1, N-2, \ldots, 1$ (backward solve)
$X_i = z_i - u_i X_{i+1}$.

computed solution at different time steps

comparison of two schemes: $k = 5.10^{-5}$

We give in an Appendix to this Chapter the Fortran program to solve the heat equation (134)-(136) on $]0,1[$, in the homogeneous case (*i.e.* for $g = 0$), using two different schemes: the explicit scheme, and the implicit scheme using the LU method as described above. The initial condition is $\varphi^0(x) = x(1_x)$.

Figure VII.7 plots the solution computed by the explicit scheme at different time steps: the initial time, on the curve with diamond symbols, then the time

steps $T = 0.01$ and $T = 0.1$, shown on the curves with square and cross symbols respectively. The space mesh has 100 nodes, the critical time step k_c for the explicit scheme is $k = 5.1\ 10^{-5}$, and we have taken $k = 10^{-6}$. For clarity's sake, we have not plotted on this figure the results obtained using the implicit scheme; let us just note that, for this latter scheme, we have obtained exactly identical results with only $k = 10^{-3}$!

We have compared, on figures VII.8 and VII.9, the solutions at time $T = 0.01$ obtained with, either the explicit scheme (curve with diamond symbols), or the implicit scheme (curve with square symbols); the time step is the same for both schemes. Figure VII.8 corresponds to a time step $k = 5.10^{-5} < k_c$, whereas $k = 6.10^{-5} > k_c$ on figure VII.9: the stability results are quite striking, as the explicit scheme "explodes" in the second case (take note that the scales on both figure are different).

comparison of two schemes: $k = 6.10^{-5}$

4.6 Two dimensional case

Let us consider, for example, the heat equation in a square $]0, L[^2$ of R^2, with a positive viscosity coefficient μ and homogeneous Neumann boundary conditions:

$$\frac{\partial \varphi}{\partial t} - \mu\left(\frac{\partial^2 \varphi}{\partial x_1^2} + \frac{\partial^2 \varphi}{\partial x_2^2}\right) = f \text{ in }]0, L[^2 \times]0, T[, \tag{165}$$

$$\varphi(x, 0) = \varphi^0(x) \text{ in }]0, L[^2, \tag{166}$$

$$\frac{\partial \varphi}{\partial n}(x, t) = 0 \text{ on } \partial(]0, L[^2) \times]0, T[. \tag{167}$$

The domain is partitioned into small elementary cells of size $h = 1/(N + 1)$ in each direction, and the Laplacian is approximated at an internal mesh point by the five points scheme (30). If we use a fully implicit scheme in time, this gives us:

$$\forall i, j \in \{1, ..., N\}, \quad \frac{1}{k}[\psi_{ij}^{n+1} - \psi_{ij}^n] - \frac{\mu}{h^2}[\psi_{i+1,j}^{n+1} - 2\psi_{i,j}^{n+1} + \psi_{i-1,j}^{n+1} \tag{168}$$

$$+\psi_{i,j+1}^{n+1} - 2\psi_{i,j}^{n+1} + \psi_{i,j-1}^{n+1}] = f_{ij}^{n+1}, \tag{169}$$

where ψ_{ij}^n is an approximation of the exact solution φ at the node (ih, jh) and at time $t = nk$.

At the initial time, we set $\forall i, j,\ \psi_{ij}^0 = \varphi^0(ih, jh)$. The Neumann condition on the boundary of the domain are discretized in the following way:

$$\forall j \in \{1, ..., N\}, \quad \psi_{0j}^n = \psi_{1j}^n, \quad \psi_{N-1,j}^n = \psi_{N,j}^n \tag{170}$$

$$\forall i \in \{1, ..., N\}, \quad \psi_{i0}^n = \psi_{i1}^n, \quad \psi_{i,N-1}^n = \psi_{iN}^n. \tag{171}$$

The N^2 equations included in (168 can be written in matrix form

$$AX = b, \tag{172}$$

where the matrix A is pentadiagonal, block tridiagonal, with each of the blocks a $N \times N$ square matrix;

$$A = \begin{bmatrix} A_1 & B & 0 & \cdots & \cdots & 0 \\ B & A_2 & B & 0 & \cdots & 0 \\ 0 & B & A_2 & B & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & B & A_2 & B \\ 0 & \cdots & \cdots & 0 & B & A_1 \end{bmatrix}, \quad B = \begin{bmatrix} b & 0 & \cdots & 0 \\ 0 & b & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & b \end{bmatrix}; \tag{173}$$

the diagonal blocks are written

$$A_1 = \begin{bmatrix} a_1 & b & 0 & \cdots & 0 \\ b & a_2 & b & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & b & a_2 & b \\ 0 & \cdots & 0 & b & a_1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_2 & b & 0 & \cdots & 0 \\ b & a_3 & b & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & b & a_3 & b \\ 0 & \cdots & 0 & b & a_2 \end{bmatrix}, \tag{174}$$

with

$$a_i = \frac{1}{k} + (1+i)\frac{\mu}{h^2}, \quad b = -\frac{\mu}{h^2}. \tag{175}$$

To solve the system (172), we could use the Choleski algorithm, or the preconditioned conjugate gradient method. The *successive over-relaxation* method, although slower, is simpler to program and does not require storing the matrix; we shall briefly recall its principle.

The Gauss-Seidel algorithm for solving a linear system $Ax = f$ of size $N \times N$ is an iterative method, where the approximate solution at step $m+1$, denoted by x^{m+1}, is computed from that at the previous step by the relation:

$$x_i^{m+1} = [f_i - \sum_{j<i} A_{ij} x_j^{m+1} - \sum_{j>i} A_{ij} x_j^m]/A_{ii}. \tag{176}$$

We note that in this computation, only the entries of the vector x^m with $j > i$ are used, and none of the previous entries, so that it is not necessary to store 2 vectors x^m and x^{m+1}: only one vector x is sufficient. As the computation proceeds, the entries of this vector, that are those of x^m at the outset, get replaced by those of the vector x^{m+1}. We can thus suppress the index m. Last, the scheme can be improved by introducing a positive relaxation parameter ω, so that the successive over-relaxation **algorithm** can schematically be written as:

loop over m

loop over i

$$x_i^* \leftarrow (f_i - \sum_{j \neq i} A_{ij} x_j)/A_{ii}), \tag{177}$$

$$x_i \leftarrow x_i + \omega(x_i^* - x_i)\ ; \tag{178}$$

end of the loops

A reasonable choice for the relaxation parameter, in the case of the heat equation, is to take ω on the order of 1.7.

4.7 Explicit Runge-Kutta schemes

4.7.1 Introduction

Among the possible choices for explicit schemes, let us point out the Runge-Kutta schemes, frequently used in the approximation of ordinary differential equations. For example, the second order method for solving $d\varphi/dt(t) = F(\varphi(t), t)$ can be written as

$$\frac{1}{k}[\psi^{n+1} - \psi^n] = F(\psi^{n+1/2}, (n+1/2)k), \quad \psi^{n+1/2} = \psi^n + \frac{k}{2}F(\psi^n, nk), \quad (179)$$

which is justified by noting that:

$$\varphi(t^{n+1/2}) = \varphi(t^n) + \frac{k}{2}F(\varphi(t^n), t^n) + O(k^2). \quad (180)$$

This scheme is second order, *i.e.* the consistency error is in $O(k^2)$.

We can apply this type of schemes to the solution of the heat equation (55)-(56), because if we discretize this equation only in the space variables, for example by writing the scheme (with obvious notation):

$$\frac{\partial \psi_i}{\partial t}(t) - \mathcal{D}_i^+ \mathcal{D}_i^- \, \psi(t) = 0, \quad (181)$$

we obtain a system of ordinary differential equations only depending on the t variable. The second order Runge-Kutta scheme (181) can then be written:

$$\frac{1}{k}[\psi_i^{n+1} - \psi_i^n] = \mathcal{D}_i^+ \mathcal{D}_i^- (\Psi^{n+1/2}), \quad \psi_i^{n+1/2} = \psi_i^n + \frac{k}{2}\mathcal{D}_i^+ \mathcal{D}_i^- (\Psi^n), \quad (182)$$

still denoting by Ψ^n the vector with entries ψ_i^n. Let us analyze the von Neumann stability of this scheme. By a Fourier transform, we obtain, setting $\lambda = (2k/h^2)\sin^2(h\xi/2) \geq 0$, and with ξ being the dual variable to the space variable x:

$$\hat{\psi}^{n+1}(\xi) = \hat{\psi}^n(\xi) - 2\lambda \hat{\psi}^{n+\frac{1}{2}}(\xi), \quad \hat{\psi}^{n+\frac{1}{2}}(\xi) = (1-\lambda)\hat{\psi}^n(\xi), \quad (183)$$

which gives eventually:

$$\hat{\psi}^{n+1}(\xi) = \hat{\psi}^n(\xi)\,[\,1 - 2\lambda(1-\lambda)\,]. \quad (184)$$

The above polynomial in λ being always positive, it will be strictly less than 1, for all ξ, if and only if the stability condition $k < h^2/2$ is satisfied.

Under this stability condition, we thus obtain a convergent scheme of second order in time and space. It is of course possible to extend this analysis to higher order Runge-Kutta schemes.

4.7.2 Runge–Kutta methods.

We recall the definition of a Runge–Kutta method for an ordinary differential equation.

Consider the ordinary differential equation

$$\frac{dy}{dt}(t) = f(t, y(t)). \tag{185}$$

The Euler scheme is defined by

$$\bar{y}^h_{k+1} = \bar{y}^h_k + hf(kh, \bar{y}^h_k). \tag{186}$$

When the function f is Lipschitz, the convergence rate of the Euler scheme is of order h: there exists a constant C such that, for all time step $h < 1$ such that T/h is an integer, there holds

$$\max_{k=0,T/h} |y(kh) - y^h_k| \leq Ch.$$

A scheme is said of order R if

$$\max_{k=0,T/h} |y(kh) - y^h_k| \leq Ch^R.$$

One can construct schemes of order $R > 1$ when the function f is of class $\mathcal{C}^m([0,T] \times R^2)$ with m large enough. For example, one can derive such schemes from Taylor formula applied to $y((k+1)h) - y(kh)$. That procedure makes appear the successive derivatives of the function f, which may lead to large computation times (generally, the computation of $\frac{\partial f}{\partial x}(x)$ requires more operations than the computation of $f(x)$), or to numerically unstable algorithms (the upper bound of $f'(x)$ may be much larger than the upper bound of $f(x)$). Runge–Kutta methods are constructed to avoid the successive derivatives of f in the discretization scheme. For example, suppose that f is of class $\mathcal{C}^3([0,t] \times R^2)$, and seek constants a_1, a_2, p_1, p_2 such that the scheme

$$\bar{y}^h_{k+1} = \bar{y}^h_k + h\left\{a_1 f(kh, \bar{y}^h_k) + a_2 f(kh + p_1 h, \bar{y}^h_k + p_2 hf(kh, \bar{y}^h_k))\right\} \tag{187}$$

is of second order. Make a Taylor expansion of $\bar{y}^h_{k+1} - \bar{y}^h_k$ make it coincide up to the order 2 with the value at time $(k+1)h$ of the solution to the differential equation with initial condition at time kh equal to $\bar{y}^h_k$:

$$\begin{aligned} \bar{y}^h_k + \int_{kh}^{(k+1)h} f(s, y(s))ds \quad &\simeq \quad \bar{y}^h_k + hf(kh, \bar{y}^h_k) \\ &\quad + \frac{1}{2}h^2 \left(\frac{\partial f}{\partial t}(kh, \bar{y}^h_k) + f(kh, \bar{y}^h_k)\frac{\partial f}{\partial y}(kh, \bar{y}^h_k)\right). \end{aligned}$$

By identification, one deduces

$$\begin{aligned} a_1 + a_2 &= 1 \\ a_2 p_1 = a_2 p_2 &= \frac{1}{2}. \end{aligned}$$

For example, one can choose $a_1 = a_2 = \frac{1}{2}$ et $p_1 = p_2 = 1$.

More generally, the principle is as follows:

$$\bar{y}_{k+1}^h = \bar{y}_k^h + h \sum_{i=0}^{Q} a_i F_i(kh, \bar{y}_k^h)$$

with

$$\begin{aligned} F_0(t,y) &= f(t,y) \\ F_1(t,y) &= f(t + p_1 h, y + p_2 h F_0(t,y)) \\ F_2 t, y) &= f(t + p_2, y + p_3 F_0(t,y) + p_4 h F_1(t,y)), \cdots \end{aligned}$$

The method of 4th order is often used in practice:

$$\begin{aligned} y_{k+1}^h &= \bar{y}_k^h + \frac{1}{6} h (F_0(kh, \bar{y}_k^h) + 2F_1(kh, \bar{y}_k^h) + 2F_2(kh, \bar{y}_k^h) && (188) \\ &\quad + F_3(kh, \bar{y}_k^h)), && (189) \end{aligned}$$

with

$$\begin{aligned} F_0(t,y) &:= f(t,y) \\ F_1(t,y) &:= f\left(t + \frac{h}{2}, y + \frac{h}{2} F_0(t,y)\right) \\ F_2(t,y) &:= f\left(t + \frac{h}{2}, y + \frac{h}{2} F_1(t,y)\right) \\ F_3(t,y) &:= f\left(t + h, y + h F_2(t,y)\right). \end{aligned}$$

We now come back to the solving of differential systems of the type

$$\frac{d}{dt} u_\delta(t, x_i) + A_\delta u_\delta(t, x_i) = 0. \tag{190}$$

In option models, the use of a Runge–Kutta method is likely to recommend since the solution is smooth. Such a method is explicit, and thus is conditionally stable. On the other hand, the order of convergence in time can be high, for example if one uses a 4th order method. Consequently, h does not need to be chosen very small. In definitive, such an explicit method may be more efficient than an implicit method which, as we have seen, requires the resolution of a linear system at each time step.

5 Finite Differences for a convection equation

We leave the framework of parabolic equations to take on that of hyperbolic equations, starting with the study of *convection* equations, that frequently occur in practice. Let Ω be a subset of R^d and let $u = (u_1, u_2, ..., u_d)$ be a given vector field defined on Ω; we seek a function φ satisfying the partial differential equation

$$\frac{\partial \varphi}{\partial t} + u \cdot \nabla \varphi = f \text{ in } \Omega \times]0, T[, \tag{191}$$

with the initial condition

$$\varphi(x,0) = \varphi^0(x), \quad \forall x \in \Omega \ ; \tag{192}$$

we recall the notation:

$$u \cdot \nabla\varphi = \sum_{i=1}^{d} u_i \frac{\partial \varphi}{\partial x_i}. \tag{193}$$

What are the possible boundary conditions for such a problem? It is clear that since the spatial derivative is only of first order, we cannot prescribe the function φ on the whole boundary Γ of Ω (to convince oneself of that fact, it suffices, in one dimension, to compute explicitly the exact solution).

For simplicity's sake, we shall assume that the velocity field u is independent of the variable t. Let us denote by n the exterior normal to Γ and denote by Γ^- the part of the boundary where the velocity field u is incoming *i.e.*:

$$\Gamma^- = \{x \in \Gamma : \quad u(x).n(x) < 0\}. \tag{194}$$

Mathematically, the problem is well posed if we take as a boundary condition:

$$\varphi(x,t) = g(x), \quad \forall x \in \Gamma^-, \quad \forall t \in]0,T[. \tag{195}$$

Remark It is possible to obtain an analytical solution of problem (191)-(195) by "integrating" the equation along the *characteristic* curves of the vector filed u. These curves ($t \to X(t)$) are defined by:

$$\frac{dX}{dt}(t) = u(X(t),t), \quad X(t=0) = x. \tag{196}$$

Show that, if we set $\Phi(t) = \varphi(X(t),t)$, we have:

$$\Phi'(t) = (\frac{\partial \varphi}{\partial t} + (u \cdot \nabla)\varphi)(X(t),t). \tag{197}$$

The function Φ is thus constant; deduce from this the expression of φ.

We shall now propose several schemes for approximating with finite differences this convection equation, starting by the one dimensional case, and assuming that the space variable varies over the whole of R.

5.1 Lax' Scheme

Let us approximate the continuous operator by $D_t^+ + uD_x^-$; we obtain the explicit backward Euler scheme defined by

$$\frac{1}{k}[\psi_i^{n+1} - \psi_i^n] + \frac{u_i}{h}[\psi_i^n - \psi_{i-1}^n] = f_i^n, \tag{198}$$

with the usual initial condition ($h = \delta x; x_i = ih$):

$$\forall i,\ \psi_i^0 = \varphi^0(x_i). \tag{199}$$

The consistency error of the scheme is in $O(k+h)$. To simplify studying stability, we shall assume that u is constant in time and space. By Fourier transform, we obtain:

$$\hat{\psi}^{n+1}(\xi) - \hat{\psi}^n(\xi) + \frac{ku}{h}\hat{\psi}^n(\xi)[1 - e^{i\xi h}] = k\hat{f}^n. \tag{200}$$

The *amplification factor* is thus the complex number defined by

$$a(\xi) = 1 - \frac{ku}{h}[1 - e^{i\xi h}], \tag{201}$$

whose modulus square is equal to:

$$a(\xi)|^2 = [1 - \frac{ku}{h}(1 - \cos\xi h)]^2 + [\frac{ku}{h}\sin\xi h]^2 \tag{202}$$

$$= 1 - 2\frac{ku}{h}(1 - \cos\xi h) + 2(\frac{ku}{h})^2(1 - \cos\xi h) \tag{203}$$

$$= 1 - 2\frac{ku}{h} + 2(\frac{ku}{h})^2 + 2\frac{ku}{h}(1 - \frac{ku}{h})\cos\xi h \tag{204}$$

$$\leq 1 \ \text{ si } \ \frac{ku}{h} \leq 1 \ \text{ and } \ u \geq 0. \tag{205}$$

The proposed scheme is thus von Neumann stable under the two conditions:

$$\frac{ku}{h} \leq 1 \ \text{ et } \ u \geq 0. \tag{206}$$

The first condition is the *CFL condition* (Courant-Friedrich-Levy). As far as the second condition is concerned, we can note that if $u < 0$, scheme (198) is *always unstable*; to have stability in that case, we must use a downwind scheme, that is:

$$\frac{1}{k}[\psi_i^{n+1} - \psi_i^n] + \frac{u_i}{h}[\psi_{i+1}^n - \psi_i^n] = f_i^n. \tag{207}$$

One can show in the same way that, if u id constant in time and space, this latter scheme is stable under the condition:

$$\frac{k|u|}{h} \leq 1 \ \text{ and } \ u \leq 0. \tag{208}$$

Remark 4.1 Let us note that the CFL condition we found is much less restrictive for the time step than the stability condition we had found for the heat equation. This is of course because the problem is here of first order in in space, whereas it was of second order for the heat equation.

Remark 4.2 The *CFL condition* obtained for the scheme (198) can be interpreted graphically in the (x,t) space, as on figure VII.10. The half-line $M^{n+1}y$ from point M^{n+1} with slope $1/u$ must intersect the line $t = t^n$ at a point of segment N^nM^n, in other words the "numerical domain of dependence", which is the surface S bounded by the triangle with vertices N^n, M^n and M^{n+1} must contain the "exact domain of dependence" *i.e.* the half-line $M^{n+1}y$. We also see that if u is negative, S does not contain this half-line: there is instability.

5.2 Lax-Wendroff scheme

The above scheme is only first order in time and space; we shall construct, using a Taylor expansion, a scheme accurate to second order. We still assume u is constant and we shall set $f = 0$ for simplicity's sake; we have, denoting the exact solution of (191)-(195) by φ,

$$\varphi(x,t+k) = \varphi(x,t) + k\frac{\partial\varphi}{\partial t}(x,t) + \frac{k^2}{2}\frac{\partial^2\varphi}{\partial t^2}(x,t) + 0(k^3). \tag{209}$$

If we express that φ is a solution of (191) with $f = 0$, we obtain

$$\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial}{\partial t}(-u\frac{\partial\varphi}{\partial x}) = -u\frac{\partial}{\partial x}(\frac{\partial\varphi}{\partial t}) \tag{210}$$

$$= -u\frac{\partial}{\partial x}(-u\frac{\partial\varphi}{\partial x}) = u^2\frac{\partial^2\varphi}{\partial x^2}, \tag{211}$$

so that

$$\varphi(x,t+k) = \varphi(x,t) - ku\frac{\partial\varphi}{\partial x}(x,t) + \frac{k^2}{2}u^2\frac{\partial^2\varphi}{\partial x^2}(x,t) + 0(k^3). \tag{212}$$

Using centered finite difference approximations of the space variables derivatives, we then obtain the Lax-Wendroff scheme:

$$\psi_i^{n+1} = \psi_i^n - ku(\frac{\psi_{i+1}^n - \psi_{i-1}^n}{2h}) + \frac{k^2}{2}\frac{u^2}{h^2}[\psi_{j+1}^n - 2\psi_i^n + \psi_{i-1}^n]. \tag{213}$$

By construction, this scheme is second order in time and space.

it is

5.3 The multidimensional case

There are no additional difficulties; in 2 dimensions, the Lax-Wendroff scheme for instance becomes (taking $f = 0$ for simplicity's sake):

$$\psi_{ij}^{n+1} = \psi_{ij}^n - \frac{ku_{ij}}{2h}(\psi_{i+1,j}^n - \psi_{i-1,j}^n) - \frac{kv_{ij}}{2h}(\psi_{i,j+1}^n - \psi_{i,j-1}^n) \tag{214}$$

$$+\frac{k^2}{2}\big[u_{ij}^2 D_{x,i}^+ D_{x,i}^+ \Psi_{.j}^n + u_{ij}v_{ij}(D_{x,i}^+ D_{y,j}^- + D_{x,i}^- D_{y,j}^+)\Psi_{..}^n \tag{215}$$

$$+v_{ij}^2 D_{y,j}^+ D_{y,j}^+ \Psi_{i.}^n\big]. \tag{216}$$

where we have denoted the discrete operators with different notation according to the x and y directions. Index i refers to the first variable x, whereas j is associated with the second variable y: for example $\Psi^n_{\cdot,j}$ denotes the singly indexed sequence $(\psi^n_{i,j})_i$ indexed by i whereas $\Psi^n_{\cdot\cdot}$ is the doubly indexed sequence $(\psi^n_{ij})_{i,j}$ indexed by i and j. Here, the velocity field has two components $\vec{u} = (u, v)$.

In this scheme, the second mixed derivative is approximated by the following 7 points scheme:

$$\frac{\partial^2\varphi}{\partial x\partial y}(ih, jh) \simeq \frac{1}{h^2}[-\varphi((i-1)h, (j+1)h) + \varphi(ih, (j+1)h) + \varphi((i-1)h, jh) \quad (217)$$

$$-2\varphi(ih, jh) + \varphi((i+1)h, jh) + \varphi(ih, (j-1)h) - \varphi((i+1)h, (j-1)h)] \quad (218)$$

Stability analysis is done via Fourier transform in both variables x and y, assuming the velocity field $\vec{u}$ is constant.

6 The convection-diffusion equation

Let us consider the following equation

$$\frac{\partial\varphi}{\partial t} - \nu\frac{\partial^2\varphi}{\partial x^2} + u\frac{\partial\varphi}{\partial x} + r\varphi = 0 \text{ in } R\times]0, T[, \quad (219)$$

$$\varphi(x, 0) = \varphi^0(x), \quad (220)$$

where ν is a real positive parameter, u is a given velocity field and r is a function defined over $R\times]0, T[$.

6.1 Continuous case

If we multiply equation (220) by φ, integrate over the whole space, and integrate by parts, we obtain, omitting the integration variables to simplify the notation:

$$\frac{1}{2}\frac{\partial}{\partial t}\Big(\int_R \varphi^2\Big) + \nu\int_\Omega \nabla|\varphi|^2 - \int_R \frac{\varphi^2}{2}\frac{\partial u}{\partial x} + \int_\Omega r\varphi^2 = 0. \quad (221)$$

The kinetic energy $E = \int_R \varphi^2$ decreases with increasing time as soon as the following condition is satisfied:

$$-\frac{1}{2}\frac{\partial u}{\partial x} + r \geq 0. \quad (222)$$

However a simple change of variables shows that this condition is not indispensable in finite time. Indeed, if we set $\varphi_1 = e^{-\alpha t}\varphi$, equation (220) becomes

$$\frac{\partial\varphi_1}{\partial t} + \alpha\varphi_1 - \nu\frac{\partial^2\varphi_1}{\partial x^2} + u\frac{\partial\varphi_1}{\partial x} + r\varphi_1 = 0, \quad (223)$$

i.e. it is of the same type as (220) except that r is changed to $r + \alpha$. It can be interesting, numerically, to carry out this change of variables, so as to avoid the exponential growth of the solution with time.

6.2 Discretization

Let us apply the following Crank-Nicolson type scheme; we obtain:

$$\frac{1}{k}[\psi_j^{n+1}-\psi_j^n]-\frac{1}{2}\frac{\nu}{h^2}[\psi_{j+1}^{n+1}-2\psi_j^{n+1}+\psi_{j-1}^{n+1}]-\frac{1}{2}\frac{\nu}{h^2}[\psi_{j+1}^{n}-2\psi_j^{n}+\psi_{j-1}^{n}]$$
$$+\frac{u_j^{n+1/2}}{2h}(\psi_{j+1}^{n+1}-\psi_j^{n+1})+\frac{u_j^{n+1/2}}{2h}(\psi_j^{n}-\psi_{j-1}^{n})$$
$$+\frac{r_j^{n+1}}{2}\psi_j^{n+1}+\frac{r_j^{n}}{2}\psi_j^{n}=0. \tag{224}$$

Let us assume, for simplicity's sake, that all coefficients are constant, and let us perform a stability analysis of this scheme by using Fourier transform in space; we obtain:

$$\hat{\psi}^{n+1}(\xi)-\hat{\psi}^{n}(\xi)+\frac{2\nu k}{h^2}\hat{\psi}^{n+1}(\xi)\sin^2\frac{\xi h}{2}+\frac{2\nu k}{h^2}\hat{\psi}^{n}(\xi)\sin^2\frac{\xi h}{2}$$
$$+\frac{ku}{2h}\hat{\psi}^{n+1}(\xi)(\cos\xi h+i\sin\xi h-1)+$$
$$\frac{-ku}{2h}\hat{\psi}^{n}(\xi)(\cos\xi h-i\sin\xi h-1)+k\frac{r}{2}(\hat{\psi}^{n+1}+\hat{\psi}^{n})(\xi)=0. \tag{225}$$

The *amplification factor* can thus be written

$$a(\xi)=\frac{1-\alpha+\gamma-i\beta}{1+\alpha+\gamma+i\beta}, \tag{226}$$

with

$$\alpha=\frac{2\nu k}{h^2}\sin^2(\frac{\xi h}{2})+k\frac{r}{2},\ \beta=\frac{ku}{2h}\sin(\xi h),\ \gamma=-\frac{ku}{h}\sin^2(\frac{\xi h}{2}). \tag{227}$$

We have

$$|a(\xi)|^2-1=\frac{-4\alpha(1+\gamma)}{(1+\alpha+\gamma)^2+\beta^2}, \tag{228}$$

and α is positive (according to (222) we assume $r\geq 0$ since u is constant). Thus, $|a(\xi)|^2-1$ has the same sign as $-(1+\gamma)$. The scheme will be stable if, for any ξ, $1+\gamma(\xi)\geq 0$, or if the following *stability condition* is satisfied:

$$\frac{ku}{h}\leq 1; \tag{229}$$

we note it is always satisfied if u is negative.

plot of function $A(t)=|a(\xi)|^2$, $t=\xi$

On figure VII.11 we have plotted the function $|a|^2$, for $k=h=0.1$, and different values of u: the solid line curve corresponds to $u=0$, that with diamond shaped symbols is the limit case $u=1$, and last that with + symbols

corresponds to $u = 10$. For clarity's sake, we have not shown the curve corresponding to $u = -10$ which indeed turns out to be located under the line with ordinate 1. We observe that the stability condition (229) is not satisfied for $u = 10$, whereas it is satisfied in the other cases: for $u = 10$, we must decrease the time step so as to obtain a stable scheme.

To compute the *order* of consistency of the scheme, it suffices to study that of the approximation of the first order term $u(\partial\varphi/\partial x)$, as the other terms, already estimated, give a consistency order of $h^2 + k^2$. We recall that this estimate has been obtained by expanding the exact solution φ in the neighborhood of the point $(jh, (n + 1/2)k)$. By 2 successive Taylor expansions, one with $n + 1$ fixed in the neighborhood of the point with index $j + 1/2$, the other at $j + 1/2$ fixed, in the neighborhood of $t^{n+1/2} = n + 1/2k$, we obtain:

$$\frac{1}{2h}[\varphi((j+1)h,(n+1)k) - \varphi(jh,(n+1)k)] = \frac{1}{2}\frac{\partial}{\partial x}\varphi((j+\frac{1}{2})h,(n+1)k)$$
$$+0(h^2) = \frac{1}{2}\frac{\partial}{\partial x}\varphi((j+\frac{1}{2})h,(n+\frac{1}{2})k) + \frac{k}{4}\frac{\partial^2}{\partial x\partial t}\varphi((j+\frac{1}{2})h,(n+\frac{1}{2})k)$$
$$+0(h^2+k^2). \tag{230}$$

In a similar way, we have:

$$\frac{1}{2h}[\varphi(jh,nk) - \varphi((j-1)h,nk)] = \frac{1}{2}\frac{\partial}{\partial x}\varphi((j-\frac{1}{2})h,nk) + 0(h^2)$$
$$= \frac{1}{2}\frac{\partial}{\partial x}\varphi((j-\frac{1}{2})h,(n+\frac{1}{2})k) - \frac{k}{4}\frac{\partial^2\varphi}{\partial x\partial t}((j-\frac{1}{2})h,(n+\frac{1}{2})k) + 0(h^2+k^2).$$

Using the following relations, for $p = n + 1/2$,

$$\frac{1}{2}\frac{\partial\varphi}{\partial x}((j+\frac{1}{2})h,pk) + \frac{1}{2}\frac{\partial\varphi}{\partial x}((j-\frac{1}{2})h,pk) = \frac{\partial\varphi}{\partial x}(jh,pk) + 0(h^2),$$
$$\frac{\partial^2\varphi}{\partial x\partial t}((j+\frac{1}{2})h,pk) - \frac{\partial\varphi}{\partial x\partial t}((j-\frac{1}{2})h,pk) = \frac{\partial\varphi}{\partial x}(jh,pk) + 0(h), \tag{231}$$

we obtain, by adding (230) and (231)

$$\tfrac{1}{2h}[\varphi((j+1)h,(n+1)k) - \varphi(jh,(n+1)k) + \varphi(jh,nk) - \varphi((j-1)h,nk)]$$
$$= \tfrac{\partial\varphi}{\partial x}(jh,(n+\tfrac{1}{2})k) + 0(h^2+hk+k^2), \tag{232}$$

which proves that the scheme (224) is in $0(h^2 + hk + k^2)$: if h and k are of the same order of magnitude, the proposed scheme is of second order in time and space.

scheme

7 Finite differences in time and finite elements in space

In two or more dimensions, as soon as the computational domain does not have its boundary piecewise parallel to the coordinate axes, the finite difference

method (in the space variables) is no longer practical: we must transform the partial differential equation into local coordinates before discretizing it, and this is a lengthy, and sometimes painful, computation. It may then be preferable to change the discretization method in space and to choose, for instance, a *finite element* or a finite volume method: this is what we do in this Section and the next.

Let us consider, for instance, the following convection-diffusion equation:

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot (u\varphi) - \nu \Delta \varphi = f \text{ in } \Omega \times]0, T[, \quad (233)$$

$$\varphi(x, 0) = \varphi^0(x) \text{ in } \Omega, \quad (234)$$

$$\varphi(x, t) = 0 \text{ on } \Gamma \times]0, T[, \quad (235)$$

where $u = u(x, t)$ is a given velocity field, and ν is a strictly positive constant.

To fix ideas, let us discretize the time derivative by using a fully implicit finite difference scheme. We then inductively define a sequence of functions $(x \to \phi^n(x))_n$ ($\phi^n \simeq \varphi(., t^n)$ only depending on the space variable x: we start with the initialization $\phi^0 = \varphi^0$, then, knowing ϕ^n, we compute ϕ^{n+1} by solving the following boundary value problem ($k = \delta t > 0$)

$$\text{find } \phi^{n+1} \text{ solving} \quad (236)$$

$$\frac{\phi^{n+1} - \phi^n}{k} + \nabla \cdot (u^{n+1}\phi^{n+1}) - \nu \Delta \phi^{n+1} = f^{n+1} \text{ in } \Omega, \quad (237)$$

$$\phi^{n+1} = 0 \text{ on } \Gamma, \quad (238)$$

with the notation: $u^{n+1} = u(., t^{n+1}), f^{n+1} = f(., t^{n+1})$.

We thus obtain a sequence of boundary value problems, to which we may apply the finite element technique as explained in the beginning of this book. To do this, we start with the continuous problem (in space) (7.2), and write its *variational formulation* as

$$\text{find } \phi^{n+1} \in H_0^1(\Omega) \text{ such that } \forall w \in H_0^1(\Omega), \text{ we have: } \mathcal{A}(\phi^{n+1}, w) = l(w), \quad (239)$$

where $\mathcal{A}$ and l are respectively the bilinear and linear forms defined over $H_0^1(\Omega)$ by:

$$\mathcal{A}(v, w) = \int_\Omega (vw)(x)dx - k \int_\Omega (u^{n+1} v \cdot \nabla w)(x)dx + \nu k \int_\Omega (\nabla v \cdot \nabla w)(x)dx,$$

$$l(w) = k \int_\Omega (f^{n+1} w)(x)dx + \int_\Omega (\phi^n w)(x)dx. \quad (240)$$

Using Green's formula in the second integral defining $\mathcal{A}$, we have:

$$\mathcal{A}(w, w) = \int_\Omega w^2(x)dx + \frac{k}{2} \int_\Omega [(\nabla \cdot u^{n+1}) w^2](x)dx + \nu k \int_\Omega |(\nabla w)(x)|^2 dx, \quad (241)$$

so that if $1 + (k/2)(\nabla \cdot u^{n+1}) \geq 0$, the bilinear form $\mathcal{A}$ is elliptic on $H_0^1(\Omega)$ and problem (7.3) admits a unique solution (let us note that the above condition is always satisfied if u is a divergence free field).

We then proceed to the discretization by covering Ω with triangles with vertices q^i; let us call Ω_h the computational domain defined as the union of all these triangles. Let $w^1, \ldots, w^{nv}$ be the usual basis functions for the P^1 finite element approximation ($w^i(q^j) = \delta_{ij}$). At every time step t^{n+1}, we are led to solve the above variational problem in the subspace V_h of $H_0^1(\Omega)$ generated by all the functions w^i associated with interior nodes of the mesh, *i.e.* the space of functions w_h continuous on Ω_h, equal to zero on the boundary of Ω_h whose restriction to each triangle is a linear function. The *discrete variational problem* is then written

$$\text{find } \phi_h^{n+1} \in V_h \text{ such that } \forall w_h \in V_h, \text{ we have: } \mathcal{A}_h(\phi_h^{n+1}, w_h) = l_h(w_h), \tag{242}$$

where $\mathcal{A}_\langle$ has the same expression as $\mathcal{A}$, except that we integrate over the computational domain Ω_h, and the linear form l_h is now defined by:

$$l(w_h) = k \int_{\Omega_h} (f^{n+1} w_h)(x) dx + \int_{\Omega_h} (\phi_h^n w_h)(x) dx. \tag{243}$$

The numerical technique is then the one that was explained in the first chapters: we expand ϕ_h^{n+1} on the basis of V_h and we express (7.5) by taking for w_h each of the basis functions. This leads us to the solution of a linear system whose solution gives us the components of ϕ_h^{n+1}.

If the solution of the continuous problem (7.1) is sufficiently smooth, which actually depends on the smoothness of the data, it is shown, for example, in Pironneau (1988) that this method is convergent, the error between the exact solution and the approximate solution being $O(h + k)$, if we denote by h the size of the mesh. More generally, the error would be of order r in space, should we have chosen an approximation with P^r finite elements.

It is of course possible to choose other discretization schemes in time; let us single out, for example, the Crank-Nicolson scheme defined by:

$$\text{find } \phi_h^{n+1} \in V_h \text{ such that } \forall w_h \in V_h, \text{ we have:} \tag{244}$$

$$\int_{\Omega_h} (\phi_h^{n+1} w_h)(x) dx - k \int_{\Omega_h} (\tfrac{\phi^{n+1}+\phi^n}{2} u^{n+1} \cdot \nabla w_h)(x) dx \tag{245}$$

$$+\nu k \int_{\Omega_h} (\nabla(\tfrac{\phi^{n+1}+\phi^n}{2}) \cdot \nabla w_h)(x) dx, \tag{246}$$

$$= k \int_{\Omega_h} (f^{n+\frac{1}{2}} w_h)(x) dx + \int_{\Omega_h} (\phi_h^n w_h)(x) dx. \tag{247}$$

The stability of this scheme is easily proven by energy qualities, such as were described in Section 3.4.1 (it suffices to take $w_h = \phi^{n+1} + \phi^n$). We can also make the linear system giving the components of ϕ^{n+1} in the basis of V_h explicit and study the eigenvalues of the corresponding matrix [Pironneau (1988)]. Let us note that on a uniform triangular mesh of a rectangular domain, the scheme is the same as the one we would obtain with a finite difference method, and this gives another way to check its stability. Last, this scheme is second order

in time, and for a sufficiently regular mesh, we can hope for a second order accuracy in space.

8 Finite differences in time and finite volumes in space

Let us again consider a convection-diffusion equation (the notation are the same as in the previous Section), with, for a change, Neumann boundary conditions:

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot (u\varphi) - \nu \Delta \varphi = f \text{ in } \Omega \times]0,T[, \tag{248}$$

$$\varphi(x,0) = \varphi^0(x) \text{ in } \Omega, \tag{249}$$

$$\frac{\partial \varphi}{\partial n}(x,t) = g(x,t) \text{ sur } \Gamma \times]0,T[. \tag{250}$$

We cover the domain Ω by quadrangles Q_k, so that $\cup_{k\in K} Q_k = \Omega$, and we assume that this mesh is *admissible*, *i.e.* the intersection between two quadrangles is either empty, or reduced to one point , or to a whole side. We integrate the partial differential equation on any one of those quadrangles and obtain, after using Green's formula,

$$\frac{\partial}{\partial t} \int_{Q_k} \varphi(x,t)dx + \int_{\partial Q_k} (u \cdot n\varphi)(x,t)d\gamma(x) + \nu \int_{\partial Q_k} (\frac{\partial \varphi}{\partial n})(x,t)d\gamma(x) \tag{251}$$

$$= \int_{Q_k} f(x,t)dx. \tag{252}$$

As above, we can discretize this equation in time, choosing an implicit, explicit, or semi-implicit scheme. If, for instance, we choose an explicit scheme, we are led to compute a sequence of functions ϕ^m defined by the recurrence relation

$$\int_{Q_k} \frac{1}{\delta t}[\phi^{m+1} - \phi^m](x)dx + \int_{\partial Q_k} (u^m \cdot n\phi^m)(x)d\gamma(x) \tag{253}$$

$$+\nu \int_{\partial Q_k} (\frac{\partial \phi^m}{\partial n})(x)d\gamma(x) = \int_{Q_k} f^m(x)dx, \tag{254}$$

and initialized by $\phi^0 = \varphi^0$.

a finite volume mesh defined by rectangles

8.1 A cell centered scheme

We shall evaluate each of the above integrals by using only the values of the unknown at the center of the cells, which are the nodes of the mesh. At step $n+1$, the unknowns of the problem are the values of ϕ^{n+1} at each of those nodes, so that there are as many unknowns as equations like (8.3).

Let us now describe how to approximate these integrals. To do this, let us use the notation of figure VII.12, and denote by q^k (resp. q^l) the center of cell Q_k (resp. Q_l). With a view to computing the line integrals, we denote by $\tilde{\phi}_{rs}$ the average of ϕ on the edge $[q_r, q_s]$, *i.e.*

$$\tilde{\phi}_{rs} = \frac{1}{|q^r - q^s|} \int_{]q^r,q^s[} \phi(x)d\gamma(x). \tag{255}$$

Using the midpoint *quadrature* formula (formula (208) from Chapter II), we have $\tilde{\phi}_{rs} \simeq \phi((q^r + q^s)/2)$. If the cells are orthogonal, the points q^k, q^l and $(q^r + q^s)/2$ all lie on a common line, and the value of ϕ at this latter point can then be approximated by the weighted average of the values taken by ϕ at the points q^k and q^l. This gives eventually:

$$\tilde{\phi}_{rs} \simeq \phi_{rs} = \phi_l \frac{|q^k - \frac{q_r+q_s}{2}|}{|q^k - q^l|} + \phi_k \frac{|q^l - \frac{q_r+q_s}{2}|}{|q^k - q^l|}. \tag{256}$$

We shall use this approximation for the average of ϕ over $[q^r, q^s]$ in all cases.

The integrals featured in (8.3) are then computed in the following way:

$$\int_{Q_k} \phi(x)dx \simeq \phi_k \text{ area } (Q_k) \tag{257}$$

$$\int_{]q^r,q^s[} \frac{\partial \phi}{\partial n}(x)d\gamma(x) \simeq \frac{\phi_l - \phi_k}{|q^k - q^l|}|q^r - q^s| \tag{258}$$

$$\int_{\partial Q_k} (u \cdot n\, \phi)(x)dx \simeq u \cdot n_{rs}\, \phi_{rs} + u \cdot n_{st}\, \phi_{st} + \cdots. \tag{259}$$

In this way, we obtain a numerical scheme for all almost orthogonal meshes. If this is not the case, we must modify the above approximate formulas. For example, to compute the normal derivative of ϕ, we can go back to the definition $\partial\phi/\partial n = \nabla\phi \cdot n$ and compute an approximation $\tilde{\nabla}\phi$ to $\nabla\phi$ by solving the system (with notation as on Figure VII.12):

$$(\tilde{\nabla}\phi).(q^r + q^s - q^u - q^t) = 2(\varphi_{rs} - \varphi_{ut}), \tag{260}$$

$$(\tilde{\nabla}\phi).(q^s + q^t - q^u - q^r) = 2(\varphi_{st} - \varphi_{ur}). \tag{261}$$

Stability and convergence results can be obtained for regular meshes by applying the method described above for finite differences. Other direct proofs exist, based on the consistency of the approximate flux (*i.e.* the line integrals featured in (8.3)) and the conservative nature of the scheme (*i.e.* if ϕ is a smooth function, the approximation of the integral over $Q_k \cap Q_l$ of $\partial\phi/\partial n$ in the equation over Q_k is the opposite to that associated with Q_l).

Although it is in some sense close to both the finite difference and the P^0 finite element method, the finite volume differs from those methods on other points. As the finite element method, it is based on a weak formulation of the partial differential equation, obtained by integrating it, but only against the function 1, and the sought solution is not expanded on a basis. The flux approximation uses a "finite difference" principle, but the resulting scheme is not consistent in the finite difference sense; consistency only plays a role at the integral formulation level, in the flux approximation.

8.2 Other possible schemes

As we have seen, the principle of the finite volume method is to integrate the equation over a cell, and then to interpolate this integral with function values at the nodes. We have seen the case of quadrangular cells, with nodes at the center of the cells, the interpolation being linear. Other choices are possible; let us quote, for instance

- the nodes are the vertices of the quadrangle Q_k and the control cells are the quadrangles obtained by splitting each quadrangle Q_k in 4 (along the medians) and joining together all the sub-quadrangles thus formed that have a common side;
- the nodes are the midpoints of the quadrangle edges and the cells are obtained by joining together all quadrangles having a common edge;
- in all of the above, we can replace the quadrangles by triangles; for example, the nodes are the vertices of the triangulation and the cell associated with a node is the polygon obtained by joining the mediatrice of the triangles sharing this vertex.

Naturally, for all these choices, it is necessary that the number of unknowns be exactly equal to the number of equations. Despite this, there can be problems. Let us consider, for example, the convection equation

$$\frac{\partial \varphi}{\partial t} + u \cdot \nabla \varphi = 0, \tag{262}$$

with: $\nabla \cdot u = 0$. Let us choose quadrangular cells, and define the nodes as the center of the cells; we have

$$\int_{Q_k} \frac{\partial \varphi}{\partial t} + \int_{\partial Q_k} u \cdot n \varphi = 0. \tag{263}$$

We cannot assume that φ is constant on each cell, because the second integral would be zero. This shows the importance of the choice of the interpolation.

These finite volume methods are comparatively recent [Jameson *et al.* (1986)]. They are conceptually simple and easy to implement. The theory is still under development, so that at the present time there are few theoretical error estimates.

These methods have first been developed for convection equations (Euler equation for fluid flows). Nicolades (1992) extended them to second order equations, after decoupling them into a system of two first order equations, according to the scheme:

$$\Delta \varphi = 0 \quad \leftrightarrow \quad u = \nabla \varphi, \quad \nabla \cdot u = 0. \tag{264}$$

Some of the equations are integrated over triangles, whereas the others are integrated over the Vorono polygons associated with the triangulation.

One of the interests of this method is that it makes it possible, contrary to finite difference methods, to treat equations with possibly discontinuous coefficients. Numerical tests show that for an operator like $-\nabla \cdot (\nu \nabla \varphi)$, where ν is a matrix with variable and discontinuous coefficients, the finite volume method gives, for a Neumann boundary condition, better results than the finite element method; on the other hand, the conclusion is reversed for a Dirichlet boundary condition. For a recent survey of this type of methods, the reader is referred to the forthcoming book by Eymard, Gallout and Herbin (1995), where a rather complete mathematical bibliography can be found

9 Finite Differences for Variational Inequalities

9.1 Definition

Some problems, like the Laplace equation with Dirichlet data, can be converted into a variational equation in $H_0^1(\Omega)$:

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega,\ u|_\Gamma = 0 \\ &\Leftrightarrow \\ \int_\Omega \nabla u \cdot \nabla w &= \int_\Omega fw,\ \forall w \in H_0^1(\Omega) \\ &\Leftrightarrow \\ \min_{u \in H_0^1(\Omega)} \int_\Omega (\frac{1}{2}|\nabla u|^2 &- fu) \end{aligned}$$

Other problem leads to a variational inequality in a Sobolev space.
For example let $\Omega \subset R^d$ and let

$$H_0^1(\Omega)^+ = \{w \in L^2(\Omega) \ : \ \nabla w \in L^2(\Omega)^d, \ \ w \geq 0\}$$

and consider

$$\min_{u \in H_0^1(\Omega)^+} \int_\Omega (\frac{1}{2}|\nabla u|^2 - fu) \tag{265}$$

9.2 Properties

Theorem

1. *Problem (265) has a unique solution.*
2. *It is equivalent to finding $u \in H_0^1(\Omega)^+$ such that*

$$\int_\Omega \nabla u \cdot \nabla w \geq \int_\Omega fw,\ \forall w \in H_0^1(\Omega)^+ \tag{266}$$

3. *In $\Omega^+ \equiv \{x \in \Omega \ : \ u(x) > 0\}$ we have*

$$\int_\Omega \nabla u \cdot \nabla w = \int_\Omega fw,\ \forall w \in H_0^1(\Omega^+)$$

Proof
The solution exists and is unique because it is the minimization of a strictly convex lower semi-continuous functional in a closed convex set.
By definition, if u is a solution, then all $\lambda \in R^+$ and all $w \in H_0^1(\Omega)$ such that $u + \lambda w \in H_0^1(\Omega)^+$ we have

$$\int_\Omega (\frac{1}{2}|\nabla u + \lambda w|^2 - f(u + \lambda w)) - \int_\Omega (\frac{1}{2}|\nabla u|^2 - fu) \geq 0$$

After division by λ and letting $\lambda \to 0$, (266) is found.
Furthermore is supp$w \subset \Omega^+$ then for small enough $\lambda's$

$$u + \lambda w \in H_0^1(\Omega)^+ \Rightarrow u - \lambda w \in H_0^1(\Omega)^+$$

Hence the converse inequality to (266) is obtained, so it must be an equality.

9.3 Discretization

The most obvious method of discretization would be to discretize by a Galerkin procedure whereby V_h approximates $H_0^1(\Omega)$ and V_h^+ approximates $H_0^1(\Omega)^+$:

$$V_h^+ = \{v \in V_h \ : \ v(x) \geq 0, \ \forall x \in \Omega\} \tag{267}$$

It is quite easy with piecewise linear continuous triangular elements because the hat functions w^i being positive all positive combination of these will have the required properties:

$$V_h^+ = \{v \ : \ v(x) = \sum_1^N v_i w^i(x), \quad v_i \geq 0\}$$

where N is the number of inner vertices in the triangulation.
Thus the minimization problem becomes

$$\min_{U \geq 0} \frac{1}{2} U^T A \, U - F \tag{268}$$

It has been shown earlier that the Finite Element method on a uniform triangulation for the Laplace operator is identical to the 5 points Finite Difference formula, except perhaps near the boundaries. So we have confidence is considering that (268) discretizes (266) with, in the case $d = 1$

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & & \dots & & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix} \tag{269}$$

and U, F the vectors made by the values of u and f at the grid points.

9.3.1 Linear Programming

Problem (268) is known as a quadratic programming problem with linear slack constraints. There are packages to solve it. Furthermore the optimality conditions are in the scope of linear programming packages and can be solved as such:
Find $U \in R^N, \ \Lambda \in R^d$ such that

$$A\ U - \Lambda = F \tag{270}$$
$$U \geq 0, \quad \Lambda \geq 0 \tag{271}$$

But owing to the special structure of the matrix A and the simplicity of the constraints, this may not be the best method.

9.3.2 Solution by Penalization

The easiest but perhaps not the best is to replace (268) by

$$\min_U \frac{1}{2} U^T A\ U - F + \frac{1}{2\epsilon} \sum (U_i^-)^2 \tag{272}$$

and to solve it either by an iterative gradient method or to solve the (non-linear) optimality conditions by a fixed point method like

$$AU^{n+1} + \frac{1}{\epsilon} I(U^{n-}) U^{n+1} = F$$

where $I(U^{n+})$ is the diagonal matrix with 1 on the ith row if U^n is negative and 0 otherwise.

9.3.3 Solution by Projection

A better way is to start from an initial guess U^0 and loop on

$$AU^{n+1/2} = F, \quad U_i^{n+1} = \max(U_i^{n+1/2}, 0) \tag{273}$$

The convergence can be obtained by invoking the convergence theorem for the projected gradient algorithm.Indeed a B-preconditioned gradient projection method with step-size ρ applied on (268) would give a sequence generated by

$$B\ U^{n+1/2} - B\ U^n = \rho(F - AU^n), \quad U_i^{n+1} = \max(U_i^{n+1/2}, 0)$$

Choosing $B = A$ and $\rho = 1$ gives back (273).
Note on the way that it might be more efficient to choose the ρ which minimize the criteria of (268).

9.3.4 Solution by Pseudo-time and Enthalpy

Time dependant procedures seek U as the limit in $t \to \infty$ of

$$\frac{\partial U}{\partial t} + AU = F$$

If A is a positive definite matrix, then the asymptotic solution indeed satisfies $AU = F$. Following Fremond, for variational inequalities we can study

$$\frac{1}{\beta(U)} \frac{\partial U}{\partial t} + AU = F$$

If $U \to \beta(U)$ is continuous, piecewise linear and

$$\begin{aligned} \beta(U) &= \epsilon \text{ if } U < -\epsilon \\ &= 1 \text{ if } U > 0 \end{aligned}$$

where $0 < \epsilon << 1$, then U tends to the solution of (268) when $t \to \infty$ and $\epsilon \to 0$.
In effect in the region $U < 0$ the time derivative dominates $AU - F$ and so U remains constant in time, meaning that if $U(x)$ reaches 0, it stays there and does not become negative.

9.4 Phase Changes and Free Boundaries

In phase change problems the modelling contains naturally an enthalpy function to account for the latent heat to tranform water into ice for instance. The parameter ϵ accounts for "mushy regions" of ice and water.
Let θ be the temperature of a system with water and ice. In the water Ω_w we have (κ is the thermal conductivity it is equal to κ_w in the water and κ_i in the ice):

$$\frac{1}{\beta(\theta)}\frac{\partial \theta}{\partial t} - \nabla \cdot (\kappa \nabla \theta) = 0$$

The mathematical limit $\epsilon \to 0$ may have three regions: water, ice at $\theta = 0$ and ice at $\theta < 0$.

$$\text{In the water } \Omega_w,\ \theta \geq 0 \text{ and } \frac{\partial \theta}{\partial t} - \kappa_w \Delta \theta = 0, \tag{274}$$

$$\text{In the ice } \Omega_i,\ \theta = 0 \text{ and the heat flux } \frac{\partial \theta}{\partial t} - \kappa_i \Delta \theta \geq 0. \tag{275}$$

$$\text{or } \theta < 0 \text{ and } \frac{\partial \theta}{\partial t} - \kappa_i \Delta \theta = 0. \tag{276}$$

At the interface water-ice, there is continuity of θ and of $\kappa \nabla \theta$. This problem can be viewed as a free boundary problem; the interface water-ice is also unknown but an additionnal boundary condition is given, namely the free boundary Σ is defined by $\theta = 0$ and the additional boundary condition is

$$\kappa_w \frac{\partial \theta}{\partial n}|_{\Sigma^+} = \kappa_w \frac{\partial \theta}{\partial n}|_{\Sigma^-}$$

where n is the normal to Σ and $+$ is on the side of the water.

Note that if the temperature in the ice never goes under 0 degree (that depends on $\theta_G amma$ then (275) can be summarized by writing

$$\min\{\frac{\partial \theta}{\partial t}(x,t) - \nabla \cdot (\kappa(\theta)\nabla\theta)(x,t)\ ,\ \theta(x,t)\} = 0 \ \ \forall x \in \Omega,\ \forall t \in (0,t). \tag{277}$$

Note also that with Dirichlet conditions $\theta = \theta_\Gamma$ on the boundary, the problem is equivalent to find $\theta \in H^1_{\theta_\Gamma}(\Omega)^+ = \{w \in H^1(\Omega), w \geq 0, w|_\Gamma = \theta_\Gamma\}$ such that

$$\int_\Omega (\frac{\partial \theta}{\partial t} w + \kappa(\theta) \nabla\theta \cdot \nabla w) \geq 0, \quad \forall w \in H^1_0(\Omega)^+ \tag{278}$$

and, finally, also equivalent to

$$\min_{\theta(\cdot,t) \in H^1_{\theta_\Gamma}(\Omega)^+} \int_\Omega (\theta \frac{\partial \theta}{\partial t} + \kappa(\theta) \nabla\theta \cdot \nabla\theta) \tag{279}$$

10 The BLACK and SCHOLES Equation

10.1 European Options

The Black & Scholes equation is used in finance to predict the price of an option on a share in the market.

Consider a share which is worth S_t dollars at time t (for instance x=38 at time t=0, i.e. now). We want to pay C dollars at time 0 to place an option which will give us the right to buy the share at time $T > 0$ for K dollars (K=40 for instance); we are not obliged to buy the share at time T. So obviously if the share is worth more than 40$ at time T we will exercise our right and if it is less we will not.

More precisely there will be a profit if $S_T > C + K$. The problem is to find C or more generally $C(x,t)$ for all x and use the result for $x = S_0, t = 0$.

Remark As x could be invested elsewhere in a "zero-risk" share at interest rate r, a more practical inequality would be $C(x,T) < x - Ke^{rT}$, but for our purpose that changes only the value of K.

10.1.1 Notations

t : time, x destined to be the price S_t of the share when it is used in conjunction with t; S_t : the price of the share follows a stochastic differential equation

$$dS_t = S_t(\mu dt + \sigma dw). \tag{280}$$

Thus
μ : average tendency of the price of the share per dollar
σ : volatility of the share
$C(x,t)$: price of an option on a share of value x and at time t.
r : risk free interest rate.

We know that

$$\begin{aligned} &C(x,T) = \varphi(x) \text{ is given by } \varphi(x) = max(x-K, 0) \\ &C(x,t) \approx x \text{ when } x \to +\infty \quad \forall t. \end{aligned} \tag{281}$$

10.1.2 Example

$$\sigma = 0.03,\ r = 0.1,$$
$$\mu = r,\ K = 40\$,\ T = 6 \text{ months} = 0.5$$

10.1.3 The Black and Scholes equation

The computation of C in the model of Black & Scholes involves the solution of the following parabolic equation with given final data:

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 C}{\partial x^2} + \mu x \frac{\partial C}{\partial x} - rC = 0 \quad \text{in} \quad R^+ \times]0, T[\tag{282}$$

It is also the expected value of $\varphi(S_t)e^{-r(T-t)}$ with S_t given by (280) and $S_T = x$.

When there are more than one share x is multidimensional $\mu \partial\varphi/\partial x$ is replaced by $\vec{\mu} \cdot \nabla\varphi$ and the Laplace operator $-\Delta$ replaces $-\partial^2 C/\partial x^2$.

Because of the singularity at $x = 0$ of the coefficients of the PDE, it contains a hidden boundary condition (the limit of the PDE at $x = 0$):

$$\frac{\partial C}{\partial t} - rC = 0 \quad at\ x = 0 \tag{283}$$

i.e. if r is constant:

$$C(0,t) = \varphi(0)e^{-\int_t^T r(\tau,0)d\tau} = \varphi(0)e^{r(t-T)} \tag{284}$$

10.1.4 Change of variable

To remove the singularity at $x = 0$ the following change of variable is proposed

$$u(y,t) = C(e^y, t) \tag{285}$$
$$\mu' = \mu - \frac{1}{2}\sigma^2 \tag{286}$$
$$\tau = T - t \tag{287}$$

Then the problem becomes

$$\frac{\partial \varphi}{\partial \tau} - \frac{1}{2}\sigma^2 \frac{\partial^2 \varphi}{\partial y^2} - \mu' \frac{\partial \varphi}{\partial y} + r\varphi = 0 \quad \text{in} \quad R \times]0, T[\tag{288}$$
$$u(y,0) = \varphi(e^y) \tag{289}$$

and when R is approximated by $]-L, +L[$ (this process is called "localization" in finance) then (281),(283) become

$$u(L,\tau) = e^L, \quad (i.e, C(x, T-t) \simeq x \text{ when } x >> 1) \tag{290}$$
$$u(-L,\tau) = \varphi(0)e^{r(\tau - T)} \tag{291}$$

10.1.5 Stability

Equation (289) integrated and multiplied by u gives

$$\int_\Omega \frac{1}{2}\frac{\partial u^2}{\partial \tau} + \int_\Omega \frac{u^2}{2}\nabla\cdot\vec{\mu}' + \int_\Omega \nabla u^T \frac{\sigma^2}{2}\nabla u + \int_\Omega (\nabla\cdot(\nabla\cdot\frac{\sigma^2}{2}))\frac{u^2}{2} + \int_\Omega r u^2 = \int_{\partial\Omega} ... \tag{292}$$

The 'kinetic energy' $E = \int_\Omega u^2$ will decay with time if

$$\nabla\cdot\mu + \nabla\cdot(\nabla\cdot\frac{\sigma^2}{2}) + r \geq 0 \tag{293}$$

because (292) gives a negative sign to $\partial E/\partial\tau$.

Remark

However a change of variable shows that this hypothesis is not absolutely essential.

Let $u_1 = e^{-\alpha\tau}u$ then (289) becomes

$$\frac{\partial u_1}{\partial\tau} + \alpha u_1 - \mu'\frac{\partial\varphi_1}{\partial y} - \frac{\sigma^2}{2}\frac{\partial^2 u_1}{\partial y^2} + r u_1 = 0 \tag{294}$$

so r is changed into $r+\alpha$

Numerically however if (293) is not verified it is a good idea to do this change of variable because it removes the exponential growth of u, something which is always difficult to capture because the "overflow" of real numbers when the result of an arithmetic operation is too large.

10.2 Discretization

An explicit scheme would be

$$\frac{1}{k}[u_j^{m+1} - u_j^m] - \frac{1}{2}\frac{\sigma_j^{m\,2}}{h^2}[u_{j+1}^m - 2\varphi_j^m + u_{j-1}^m] - \frac{\mu_j'^m}{h}(u_{j+1}^m - u_j^m) + \frac{r_j^m}{2}u_j^m = 0 \tag{295}$$

One of the best implicit scheme is the Crank-Nicolson scheme :

$$\begin{aligned}
&\frac{1}{k}[u_j^m - u_j^{m-1}] \\
&-\frac{1}{4}\frac{\sigma_j^{m\,2}}{h^2}[u_{j+1}^m - 2\varphi_j^m + u_{j-1}^m] - \frac{1}{4}\frac{\sigma_i^{m-1\,2}}{h^2}[u_j^{m-1} - 2\varphi_j^{m-1} + u_{j-1}^{m-1}] \\
&-\frac{\mu_j'^m}{2h}(u_{j+1}^m - u_j^m) - \frac{\mu_j'^{m-1}}{2h}(u_j^{m-1} - u_{j-1}^{m-1}) + \frac{r_j^m}{2}u_j^m + \frac{r_j^{m-1}}{2}u_j^{m-1} = 0
\end{aligned}$$

These can be applied either to the system written in (x,t) or to the one written in (y,τ)

In the first case the coefficients σ and μ are constant and the analysis of Von Neumann shows unconditional stability for the Crank-Nicolson scheme and

of course precision in $0(h^2+k^2)$. If $(x,t$ are used then the same result holds with non uniform spatial mesh obtained by transformation of the y-mesh , in principle; however in practice it is also stable for a uniform mesh in x.

11 American Options

The model now requires that $C(x,\tau)$ never becomes larger (or smaller) than $\psi(x,\tau)$ given. Thus it is a time dependant variational inequality

The problem is:

$$\max\{\frac{\partial C}{\partial \tau}+\vec{\mu}'\nabla C+\sigma^T\sigma:\nabla\nabla C-rC;\ \psi-C\}=0 \tag{296}$$

where $\sigma^T\sigma:\nabla\nabla C$ means $\sigma_{ki}\sigma_{il}\partial^2_{x_k,x_l}C$ with a summation over repeated indices.

11.1 Discretization and well posedness

Assume $\mu_i'>0$ and $d=2$. Denote by D_l^0 the central differencing operator which uses the two points left and right if l=1, or up and down if l=2, of the current point, and by $D_j^{\pm}$ the one which uses the point left (with minus sign) or right (or up or down) and the current point. Then a simple discretization of the PDE is

$$\frac{1}{k}[C_{ij}^{m+1}-C_{ij}^m]+\sum_{l=1,2}\mu_l'D_l^+C_{ij}^m+\sum_{l,n,m=1,2}\sigma_{lm}\sigma_{nm}D_l^0D_n^0C_{ij}^m-r_{ij}C_{ij}^m=0 \tag{297}$$

With boundary conditions it is a linear system for C^m: $AC^m=d$.
So to discretize (296) we consider:

$$\begin{aligned}\max\{\quad &\frac{1}{k}[C_{ij}^{m+1}-C_{ij}^m]+\sum_{l=1,2}\mu_l'D_l^+C_{ij}^m+\\ &\sum_{l,n,m=1,2}\sigma_{lm}\sigma_{nm}D_l^0D_n^0C_{ij}^m-r_{ij}C_{ij}^m\ ,\ \psi_{ij}-C_{ij}^m\}=0\end{aligned} \tag{298}$$

It is a problem of the type

$$A\ C\le d,\quad C\ge\psi\quad (C-\psi)^T(AC-d)=0 \tag{299}$$

Proposition *If δt is small enough (298) has a unique solution.*

Proof Let C_1 C_2 be two solutions, then

$$A(C_1 - C_2) = 0 \tag{300}$$
$$\text{or} \quad C_1 - C_2 = 0 \tag{301}$$
$$\text{or} \quad C_1 \geq \psi \quad AC_1 = d \quad A\, C_2 \leq d \quad C_2 = \psi \tag{302}$$
$$\text{or} \quad AC_2 = d \quad C_2 \geq \psi \quad C_1 = \psi \quad AC_1 \leq d \tag{303}$$

Therefore

$$A(C_2 - C_1) \geq 0, \quad C_2 - C_1 \geq 0 \tag{304}$$
$$\text{or} \quad A(C_2 - C_1) \leq 0, \quad C_2 - C_1 \leq 0 \tag{305}$$

In all cases $(C_2 - C_1)^T A(C_2 - C_1) \leq 0$, but $\delta t \to 0$ implies $A \to -\frac{1}{\delta t} I$, so $\|C_2 - C_1\|^2 \to \frac{\alpha}{\delta t}$ with $\alpha < 0$; therefore $C_2 = C_1$.

Remark

If $\mu = 0$ and σ is constant, then uniqueness is true for all δt.

11.1.1 Solution by truncation

Solve at each time step

$$AC^{m+1/2} = d \tag{306}$$

Set

$$C_{ij}^{m+1} = \max(C_{ij}^{m+1/2}, \psi_{ij}) \tag{307}$$

11.1.2 Solution by penalization

Consider a perturbation of (298) with $\varepsilon << 1$. Then one generates the sequence:

$$(A + \frac{1}{\varepsilon} I^m) C^{m+1} = d + \frac{1}{\varepsilon} I^m \psi \tag{308}$$

where I^m is diagonal and has ones where $C^m \leq \psi$.

11.2 Mesh refinement

The precision of the method depends much more on the mesh size here because of the change of equation at a mesh point. For precise results one ought to use variable mesh sizes; but then the computer program becomes involved and a Finite Volume Method may be prefered.

12 REFERENCE

1. R. Jarrow, A. Rudd: *Option pricing* R. Irwin publishing co. Illinois 1983.
2. Cox, Rubinstein: *Option Market* 1985.

13 Appendix A

13.1 Solution of the problem by the BPX method

Condition number of the Laplacian matrix

Scheme (30) for problem (28)-(29) leads to a linear system $A\Psi_h = F_h$. One could, of course, use a direct method (for example Choleski), but for large scale problems we would rather use an iterative method like the Conjugate Gradient. Unfortunately, the matrix A is ill conditioned and so, as soon as the number of unknowns is large (say larger than 5000) we must use a preconditioner. The *multigrid* method is particularly well suited to finite differences since constructing nested grids is of course very simple.

Let us first show that the matrix A has a condition number in $O(h^{-2})$. For simplicity's sake, let us assume that

$$h_1 = h_2 \equiv h, \quad u = 0, \quad L_1 = L_2 = 2\pi. \tag{309}$$

The scheme may be written as

$$(A\Psi_h)_{ij} \equiv -\frac{1}{h^2}(\psi_{i+1,j} + \psi_{i-1,j} + \psi_{i,j+1} + \psi_{i,j-1} - 4\psi_{i,j}) = f_{i,j} \tag{310}$$

where $\psi_{i,j} \approx u(ih, jh)$ and $i, j = 1, ..., N-1$ with $Nh = 2\pi$.

We look for the solution of the scheme in the form

$$\psi(x,y) = \sum_{\delta t,l=1,\dots,N-1} Re(u^{\delta t,l} e^{\mathbf{i}(kx+ly)}). \tag{311}$$

By substituting $u^{\delta t,l} e^{\mathbf{i}(kx+ly)}$ in the scheme, we obtain:

$$-\frac{1}{h^2} u^{\delta t,l} e^{\mathbf{i}(kih+ljh)} [e^{\mathbf{i}kh} + e^{-\mathbf{i}kh} + e^{\mathbf{i}lh} + e^{-\mathbf{i}lh} - 4]) = f_{i,j} \tag{312}$$

So, if $\hat{f}^{\delta t,l}$ is the discrete Fourier transform of f, that is

$$f(x,y) = \sum_{\delta t,l=1,\dots,N-1} Re(\hat{f}^{\delta t,l} e^{\mathbf{i}(kx+ly)}). \tag{313}$$

then $u^{\delta t,l} = \hat{f}^{\delta t,l} h^2 / (4\sin^2(kh/2) + 4\sin^2(lh/2))$ is the solution of the scheme, because

$$e^{\mathbf{i}kh} + e^{-\mathbf{i}kh} + e^{\mathbf{i}lh} + e^{-\mathbf{i}lh} = 2\cos(kh) + 2\cos(lh) = 4 - 4\sin^2\frac{kh}{2} - 4\sin^2\frac{lh}{2}. \tag{314}$$

For ψ to satisfy the boundary conditions, we must take a linear combination of functions that give only sine functions:

$$\psi(x,y) = \sum_{\delta t,l=0,\dots,N} u^{\delta t,l} \sin(kx)\sin(ly) \tag{315}$$

and this is only possible if f itself has the same form, that is if $f(x,y) = -f(2\pi - x, y) = -f(x, 2\pi - y)$.

This computation also shows that $v^{\delta t,l}(x,y) = \sin(kx)\sin(ly)$ is an eigenvector of the scheme for the eigenvalue $h^{-2}(4\sin^2(\frac{kh}{2}) + 4\sin^2(\frac{lh}{2}))$ because

$$(Av^{\delta t,l})_{i,j} = \frac{1}{h^2}(4\sin^2(kh/2) + 4\sin^2(lh/2))v^{\delta t,l}(ih, jh). \tag{316}$$

Since we have $(N-1)^2$ solutions of this form and the matrix A has size $(N-1)^2 \times (N-1)^2$ we have all the eigenvalues. The smallest one corresponds to $\delta t = l = 1$ and the largest one to $\delta t = l = \pi/h - 1$

The condition number of A, *i.e.* the ratio of its largest to its smallest eigenvalue is now

$$cond(A) = \frac{\sin^2(\frac{\pi}{2} - \frac{h}{2})}{\sin^2 \frac{h}{2}} \approx \frac{4}{h^2} \tag{317}$$

The BPX preconditioning Let us consider a *multigrid* mesh with K levels obtained by subdividing each rectangle of the initial finite difference grid into four identical rectangles. At level n the average size of the rectangles is denoted by δt_h.

We solve the system $A\Psi_h = F_h$ by the preconditioned conjugate gradient method with a preconditioning matrix C. The algorithm (20)-(29) from Chapter 4 requires us to be able to compute $C^{-1}b$ for a given vector b. Let us denote by $b_h^{\delta t}$ the P^1 function (*i.e.* piecewise linear and continuous) equal to $b_{i,j}$ at the vertex i, j of the triangular mesh obtained by subdividing each finite difference rectangle into two triangles by a diagonal, always the same (*cf.* Chapter 2, figure 2.9). We set

$$C^{-1}b = \sum_{\delta t=1}^{K} \sum_{i=1}^{N_{\delta t}} \frac{\int_\Omega b_h^{\delta t} w_i^{\delta t}\, dx}{\int_\Omega \nabla w_i^{\delta t} \cdot \nabla w_i^{\delta t}\, dx} w_{\delta t} \tag{318}$$

where the sum is over all basis functions (hat functions $w_i^{\delta t}$) of level δt, then over all K levels. Bramble et al. (1990) have shown that $C^{-1}A$ has a condition number in $O(1)$.

One should handle the coarsest level in a particular way, by solving the problem exactly. The right formula is then (the coarse level is now level 0)

$$C^{-1}b = A_0^{-1}Q_0 b + \sum_{\delta t=1}^{M} \sum_{i=1}^{N_{\delta t}} \frac{\int_\Omega b_h^{\delta t} w_i^{\delta t}\, dx}{\int_\Omega \nabla w_i^{\delta t} \cdot \nabla w_i^{\delta t}\, dx} w_{\delta t} \tag{319}$$

where A_0 is the restriction of A to the coarse basis functions and Q_0 is the projection L^2 on the coarse space.

Remark The method can also be applied with a finite element discretization as soon as we have a sequence of nested grids. The result on the condition number is independent of the space dimension.