Chapter 3

Finite differences for evolution equations

3.1 Approximation of the heat equation

We consider the following problem:

$$\begin{cases} \partial_t u - \mu \partial_{xx} u = f(x,t) \qquad x \in]0,1[,\, 0 < t \leq T, \\ u(0,t) = 0, \quad u(1,t) = 0, \qquad 0 \leq t \leq T, \\ u(x,0) = u_0(x), \qquad x \in]0,1[, \end{cases} \tag{3.1}$$

which is called the heat equation, or diffusion equation. The data here consists of the function f and the initial condition u_0, and μ is a given constant. We assume that the initial condition and the boundary conditions are compatible, meaning that the initial condition satisfies the boundary conditions $u_0(0) = u_0(1) = 0$. We do not seek to solve this PDE in a general setting (with general data, for example time-dependent boundary conditions, i.e. such that $u(0,t) = \alpha(t), u(1,t) = \beta(t)$) but rather we wish to understand how to construct a finite difference method in order to approximate the solution. We will focus on a particular case where a solution can be found relatively easily. In order to be able to perform the numerical analysis of a method, it is useful to know that the solution of the continuous problem exits in a certain space and is unique. In the case $f = 0$, we can actually show the following existence and uniqueness result (which uses Fourier series expansions).

Proposition 3.1.1 *Let $u_0 \in \mathcal{C}^2([0,1])$ with $u_0(0) = u_0(1) = 0$. Then problem (3.1) with $f = 0$ has a solution $u \in \mathcal{C}^0([0,1] \times [0,T]) \cap \mathcal{C}^1(]0,1[\times]0,T])$, $\partial_{xx} u \in \mathcal{C}^0([0,1]\times]0,T])$ and any solution of this regularity is unique.*

For the finite difference approximation to the solution to (3.1) we introduce, in addition to the uniform grid in space, a temporal mesh (a grid in time): Let $M > 0$ be an integer and take $\Delta t = \frac{T}{M}$, where we call Δt the time step and we let $t_n = n\Delta t$, $0 \leq n \leq M$. Analogously to Δt, we will use the following notation

$$\Delta x = h = \frac{1}{N+1}, \quad x_i = ih = i\Delta x, \quad i = 0, \ldots, N+1.$$

The points of the space-time grid are thus (x_j, t_n), $0 \leq j \leq N+1, 0 \leq n \leq M$. We now seek to compute values u_j^n which approximate the exact values $u(x_j, t_n)$ at those grid points using a finite difference scheme. We keep the same discretisation for the "Laplacian" $\partial_{xx}u$ as before, but depending on the approximation used for the time derivative, we obtain different methods for the equation. The first method is the following:

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} - \mu \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f_j^n, \qquad 1 \le j \le N, \quad 0 \le n \le M, \\ u_0^n = 0, \quad u_{N+1}^n = 0, \qquad 0 \le n \le M, \\ u_j^0 = u_0(x_j), \qquad 0 \le j \le N+1, \end{cases} \tag{3.2}$$

where $f_j^n = f(x_j, t_n)$. This is obtained by taking the exact equation at the point (x_j, t_n), then discretising the Laplacian by finite differences as we saw at the beginning of this chapter and finally replacing the time-derivative by a forward finite difference approximation of the form

$$\partial_t u(x_j, t_n) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t}.$$

This method is explicit in time: If we take

$$r = \mu \frac{\Delta t}{\Delta x^2}, \tag{3.3}$$

the difference equation of method (3.2) takes the form

$$u_j^{n+1} = (1-2r)u_j^n + r(u_{j+1}^n + u_{j-1}^n) + \Delta t f_j^n, \tag{3.4}$$

which is an *explicit* formula that allows us to compute u_j^{n+1} from the values $u_{j+1}^n, u_j^n, u_{j-1}^n$ which are known at time t_n, without the need to invert any function or linear system.

Let us now take for $0 \le n \le M$, $U_h^n = (u_1^n, u_2^n, ..., u_N^n)^T \in \mathbb{R}^N$, $F_h^n = (f_1^n, f_2^n, ..., f_N^n)^T \in \mathbb{R}^N$ and define the tridiagonal matrix

$$Q_1 = Q_1(r) = \begin{pmatrix} 1-2r & r & 0 & \dots & \dots & 0 \\ r & 1-2r & r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & r & 1-2r & r \\ 0 & \dots & \dots & 0 & r & 1-2r \end{pmatrix}.$$

Then the formula (3.4) written for $j = 1, ..., N$ yields the relation

$$U_h^{n+1} = Q_1 U_h^n + \Delta t F_h^n,$$

and allows us to compute the vectors U_h^n for $0 < n \le M$ starting from initial datum which is given exactly, specifically $U_h^0 = \overline{U}_h^0$ a vector in $\mathbb{R}^N$ with entries $(u_0(x_j))_{j=1,...,N}$.

If on the other hand we were to consider a backward finite difference approximation to the temporal derivative (which we write here for time t_{n+1})

$$\partial_t u(x_j, t_{n+1}) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t},$$

we would find from the exact equation at (x_j,t_{n+1}), the second method

$$\left\{\begin{array}{l} \dfrac{u_j^{n+1}-u_j^n}{\Delta t}-\mu\dfrac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{\Delta x^2}=f_j^{n+1}, \quad 1\le j\le N, \qquad 0\le n\le M, \\ u_0^n=0, \quad u_{N+1}^n=0, \qquad 0\le n\le M \\ \\ u_j^0=u_0(x_j), \qquad 0\le j\le N+1, \end{array}\right. \tag{3.5}$$

which is an *implicit* scheme. We can write the finite difference equation corresponding to method (3.5) using the notation (3.3) in the following form

$$(1+2r)u_j^{n+1}-r(u_{j+1}^{n+1}+u_{j-1}^{n+1})=u_j^n+\Delta t f_j^{n+1}, 1\le j\le N, \tag{3.6}$$

and we thus arrive at a tridiagonal matrix system which needs to be solved in order to find the N components of the vector U_h^{n+1}. We define the following $N\times N$-matrix

$$Q_2=Q_2(r)=\begin{pmatrix} 1+2r & -r & 0 & \dots & \dots & 0 \\ -r & 1+2r & -r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -r & 1+2r & -r \\ 0 & \dots & \dots & 0 & -r & 1+2r \end{pmatrix}.$$

The formula (3.6) written for $j=1,....,N$ thus yields the following linear system

$$Q_2(r)U_h^{n+1}=U_h^n+\Delta t F_h^{n+1}.$$

We can also take a centered difference approximation

$$\partial_t u(x_j,t_{n+1/2})\sim\frac{u(x_j,t_{n+1})-u(x_j,t_n)}{\Delta t}.$$

Considering the exact equation at $(x_j,t_{n+1/2})$ and then approximating the values $u(x_j,t_{n+1/2})$ in the expression

$$\partial_{xx}u(x_j,t_{n+1/2})\sim\frac{u(x_{j+1},t_{n+1/2})-2u(x_j,t_{n+1/2})+u(x_{j-1},t_{n+1/2})}{\Delta x^2}$$

using the averages $u(x_j,t_{n+1/2})\sim\frac{1}{2}(u(x(j,t_n)+u(x_j,t_{n+1}))$ we obtain a third method, given by

$$\left\{\begin{array}{l} \dfrac{u_j^{n+1}-u_j^n}{\Delta t}-\mu\dfrac{(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1})+(u_{j+1}^n-2u_j^n+u_{j-1}^n)}{2\Delta x^2}=f_j^{n+1/2}, \quad 1\le j\le N, \quad 0\le n\le M \\ u_0^n=0, \qquad u_{N+1}^n=0, \quad 0\le n\le M \\ \\ u_j^0=u_0(x_j), \qquad 0\le j\le N+1, \end{array}\right. \tag{3.7}$$

where $f_j^{n+1/2}=f(x_j,(n+1/2)\Delta t)$ or, alternatively, we may also take the approximation $\frac{1}{2}(f_j^n+f_j^{n+1})$. This method (called the Crank–Nicolson method) is also implicit. Introducing two more $N\times N$ tridiagonal matrices, say Q_3,Q_4 (exercise), we obtain a relation of the form

$$Q_3U_h^{n+1}=Q_4U_h^n+\Delta t F_h^{n+1/2}.$$

Remark 3.1.1 *In these three examples, we can check that the numerical methods correspond to having used the Euler method (forward or backward respectively) or the Crank–Nicolson method for the ODE*

$$\begin{cases} \dfrac{dU_h}{dt}(t)+\mu A_{0h}U_h(t) = f_h(t) \quad 0 < t \leq T \\ \\ U_h(0) = \overline{U}_h^0 \end{cases} \tag{3.8}$$

where A_{0h} is defined by (2.7), the N components of $U_h(t)$ approximate the values $u(x_i,t)$ and $\overline{U}_h^0$ is the vector with components $u_0(x_j)$. This system of differential equations (3.8) arises from the semi-discretisation in space by the method of finite differences as introduced in section 2.1, where the time t is being considered as a "parameter". This process of discretisation is also called the "method of lines". We then discretise in time, by computing a vector U_h^n which approximates the vector $U_h(t_n)$.

Remark 3.1.2 *We could also start by semi-discretising the problem in time, x being considered as a "parameter". For example if we use the implicit Euler method, we obtain*

$$\begin{cases} \dfrac{u^{n+1}(x)-u^n(x)}{\Delta t} - \mu\dfrac{d^2u^{n+1}}{dx^2}(x) = f^{n+1}(x) \quad x \in]0,1[, 0 \leq n \leq M-1 \\ u^n(0) = u^n(1) = 0, \quad 0 \leq n \leq M \\ u^0(x) = u_0(x), \quad x \in]0,1[, \end{cases} \tag{3.9}$$

where $f^n(x) = f(x,t_n)$ and the function $u^n(x)$ approximates $u(x,t_n)$. We then use the method of finite differences in space for (3.9), which consist of taking the equations at points x_j, and replacing the exact derivatives by finite differences as in section 2.1. In this way we obtain method 2 (3.5) where U_h^n approximates the coordinate vector $u^n(x_i)$.

These remarks will help us to understand how we can link the order in time and space of the method resulting from the two discretizations, to the respective orders of each semi-discretized method. This also allows us to understand how we may generalise the process by using other discretisation methods in time (for example a Runge-Kutta method) or in space (for example a finite element method).

For methods 2 and 3 we can check that the matrices Q_2 and Q_3 are M-matrices and thus invertible. These three aforementioned methods can therefore all be put in the general form

$$U_h^{n+1} = QU_h^n + \Delta t G_h^n, \ 0 \leq n \leq M-1, \tag{3.10}$$

where $Q = Q(r)$ is a known $N \times N$-matrix, and G_h^n is a vector in $\mathbb{R}^N$ known from the inital conditions of the problem, and $M\Delta t \leq T$. The initial condition is $U_h^0 = u_h^0 = (u_0(x_1), \ldots, u_0(x_N))^T$. This is a method with two stages in time (only the values at t_n are used for the calculation at time t_{n+1}), we also call this a single-step method in time.

Note that the boundary conditions in (3.1) are not zero, but rather are given by $u(0,t) = \alpha(t), u(1,t) = \beta(t)$. Thus these values, once discretised to $(\alpha(t_n), \beta(t_n))$ will appear in the components of the right hand side G_h^n.

Let us introduce for these methods the notions of consistency, stability and convergence. In order to do so, we consider the exact values $u(x_j,t_n)$ at the grid points and the associated vector

$$\overline{U}_h^n = (u(x_1,t_n),u(x_2,t_n),...,u(x_N,t_n))^T \in \mathbb{R}^N.$$

Then $\overline{U}_h^n$ does not satisfy the equations of the numerical method exactly (otherwise we would know how to calculate the exact solution at any point) but we may write

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t G_h^n + \Delta t K_h^n,$$

which defines the vector K_h^n, i.e.

$$K_h^n = \frac{1}{\Delta t}(\overline{U}_h^{n+1} - Q\overline{U}_h^n - \Delta t G_h^n).$$

K_h^n is called the *consistency error* at time t_n. Finally, we denote by $\|\ .\|$ a vector norm on $\mathbb{R}^N$ with an associated matrix norm $\|.\|$.

Definition 3.1.1 *We say that the method (3.10) is*
i) consistent with the PDE (3.1) if for every solution u of (3.1) we have

$$\sup_{m;m\Delta t\le T} \|K_h^m\| \to 0 \ \textit{as}\ \Delta t,\Delta x \to 0.$$

ii) The method is of order (p,q) (p in space and q in time) if, for every sufficiently regular solution u to (3.1), there is a constant $C = C(u) > 0$ such that

$$\sup_{m;m\Delta t\le T} \|K_h^m\| \le C(\Delta x^p + \Delta t^q).$$

iii) The method is stable (in the norm $\|\ .\|$) if there is a constant C_0 (which may depend on T) such that

$$\sup_{m;m\Delta t\le T} \|Q^m\| \le C_0.$$

iv) The method is convergent if

$$\sup_{m;m\Delta t\le T} \|\overline{U}_h^m - U_h^m\| \to 0 \ \textit{as}\ \Delta t,\Delta x \to 0.$$

iv) The method is convergent of order (p,q) (p in space and q in time) if, for every sufficiently regular solution u to (3.1), there is a constant $C = C(u) > 0$ such that

$$\sup_{m;m\Delta t\le T} \|\overline{U}_h^m - U_h^m\| \le C(\Delta x^p + \Delta t^q).$$

Remarks. i) Be careful with *notation*: in K_h^n, U_h^n, which are vectors of $\mathbb{R}^N$, n is an index in time; whereas in Q^m, where $Q = Q(r)$ is an $N \times N$-matrix, for integers m this means the power $Q^2 = QQ$, ... If we use a method with variable step size, in which $\Delta t_n \equiv t_{n+1} - t_n$ is not necessarily constant, then the matrix Q in (3.10) depends on n, i.e. $Q^{(n)}$, and we would then have a product $Q^{(m)}Q^{(m-1)}...Q^{(1)}$ instead of the power Q^m.

ii) The regularity required for consistency is that given in proposition 3.1.1; to estimate the order however, we must assume that the solution is more regular in order to be able to be able to consider Taylor series expansions (cf. proposition 3.1.2 below). Although no general result of the regularity of the solution in terms of the data has been given here (which would be the analogue of theorem 1.4.3 in the case of the heat equation) this regularity is actually obtained and can be derived from the regularity assumptions on the data.

iii) Even if we did not mention this explicitly the norm used on $\mathbb{R}^N$ must be such that the quantities involved continue to make sense when $N \to \infty$, for example the ℓ^∞-norm or the discrete ℓ^2-norm $\|\cdot\|_{2,\Delta}$ which we introduced in definition 2.2.3, which can be written here as

$$\|U_h^j\|_{2,\Delta}^2 = h \sum_{j=1}^{N} |u_j^n|^2,$$

because the values at the end points, u_0^n and u_{N+1}^n, are zero. Note that the factor h changes the value of the norm, but not the associated matrix norm. Furthermore the norm $\|\cdot\|_{2,\Delta}$ is bounded above by the norm $\|\cdot\|_\infty$. □

The following is a fundamental result sometimes called the *Lax equivalence theorem*, which relates the concepts of consistency, stability and convergence.

Theorem 3.1.1 *If a method is consistent and stable then it is convergent.*

Proof. We introduce the error $E_h^n = \overline{U}_h^n - U_h^n, 0 \le n \le M = T/\Delta t$. Then $e_h^0 = 0$ by our choice of initial condition, and we obtain from the difference of the following two relations

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t G_h^n + \Delta t K_h^n, \quad U_h^{n+1} = QU_h^n + \Delta t G_h^n$$

that, for every $n \le M-1$,

$$E_h^{n+1} = QE_h^n + \Delta t K_h^n.$$

Iterating this process we obtain, since $E_h^0 = 0$,

$$E_h^{n+1} = \Delta t \sum_{k=0}^{n} Q^k K_h^{n-k}.$$

We thus deduce that

$$\|E_h^{n+1}\| \le \Delta t \sum_{k=0}^{n} \|Q^k\| \, \|K_h^{n-k}\|.$$

If the method is stable we have

$$\|E_h^{n+1}\| \le C_0(n+1)\Delta t \sup_{0 \le k \le n} \|K_h^{n-k}\|,$$

and therefore, because this holds for every $n \le M-1$, with $M\Delta t = T$ we have

$$\sup_{m; m\Delta t \le T} \|E_h^m\| \le C_0 T \sup_{n; n\Delta t \le T} \|K_h^n\|,$$

and if the method is consistent the right hand side tends to 0 as $\Delta t, \Delta x \to 0$. □

Remark 3.1.3 *i) Using the calculation from the above proof we can show that if we have a perturbation P_h^0 in the initial data, if the method computes a sequence of values $\tilde{u}_h^n$ which are the solution to a perturbed method*

$$\tilde{U}_h^{n+1} = Q\tilde{U}_h^n + \Delta t G_h^n + \Delta t P_h^n,$$

with $\tilde{U}_h^0 = U_h^0 + P_h^0$, and if the method is stable, then the difference $\|\tilde{U}_h^n - U_h^n\|$ remains bounded in terms of the perturbations $\|P_h^0\|$ and $\max_{n\leq M}\|P_h^n\|$. This observation explains the term stability.

ii) The principle "stability + consistency implies convergence" applies very generally to a range of discretisation methodologies.

iii) Indeed the above proof also shows that if a method is consistent of order order (p,q) (p in space and q in time) and stable, then it is also convergent at that same order (p,q).

Let us now apply this result to the methods 1 and 2. We must therefore check the consistency and the stability of the methods; let us begin with studying the explicit **method 1**. For this method the consistency error is given by the equation

$$u(x_j,t_{n+1}) = (1-2r)u(x_j,t_n) + r(u(x_{j+1},t_n) + u(x_{j-1},t_n)) + \Delta t f_j^n + \Delta t \tau_j^n.$$

Proposition 3.1.2 *Assume that the solution u of problem (3.1) satisfies: $u \in \mathcal{C}^0([0,1]\times[0,T] \cap \mathcal{C}^1(]0,1[\times]0,T])$, and $\frac{\partial^4 u}{\partial x^4}, \frac{\partial^2 u}{\partial t^2} \in \mathcal{C}^0([0,1]\times]0,T])$. Then, for the method (3.2), the consistency error satisfies*

$$\sup_{m;m\Delta t\leq T} \|K_h^m\|_\infty \leq C(\Delta x^2 + \Delta t),$$

where the constant C depends on $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^4 u}{\partial x^4}|$, $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^2 u}{\partial t^2}|$.

Proof. The consistency error κ_j^n satisfies

$$\kappa_j^n = \frac{1}{\Delta t}\big(u(x_j,t_{n+1}) - u(x_j,t_n)\big) - \mu\frac{1}{\Delta x^2}\Big(u(x_{j+1},t_n) - 2u(x_j,t_n) + u(x_{j-1},t_n)\Big) - f(x_j,t_n).$$

We use again Taylor's theorem. For a function ϕ which is assumed to be of class $\mathcal{C}^2$ on $[0,T]$, we can write: for every $n \in \{0,\ldots,M-1\}$, there is a real $\theta^{(n)} \in]0,1[$ such that

$$\phi(t_{n+1}) = \phi(t_n) + \Delta t\phi'(t_n) + \frac{\Delta t^2}{2}\phi''(t_n + \theta^{(n)}\Delta t).$$

We apply this result to $\phi(t) = u(x_j,t)$, where $\phi'(t_n) = \partial_t u(x_j,t_n)$. Similarly, we can apply theorem 2.1.1 to $\varphi(x) = u(x,t_n)$ where $\varphi''(x_j) = \partial_{xx}u(x_j,t_n)$. Since u is a solution to (3.1), we have

$$\partial_t u(x_j,t_n) - \mu\partial_{xx}u(x_j,t_n) = f(x_j,t_n)$$

thus

$$\kappa_j^n = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x_j, t_n + \theta^{(n)}\Delta t) - \mu\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j + \theta_j\Delta_j, t_n),$$

for some $\theta_j, \theta^{(n)} \in]-1,1[$, whence the result follows. We have thus shown that the method is of order 2 in space and 1 in time, as was expected since the method was constructed from a second-order finite difference scheme in space (associated with the discrete Laplacian) and a first-order finite difference scheme in time (the Euler method). □

We note that, since $\|\cdot\|_\infty$ bounds $\|\cdot\|_{2,\Delta}$ from above, we also have

$$\sup_{m;m\Delta t\le T} \|K_h^m\|_{2,\Delta} \le C(\Delta x^2 + \Delta t)$$

For the stability, we have the following result.

Proposition 3.1.3 *Assume* $0 < r \le 1/2$. *Then the method is stable in the norms* $\|.\|_\infty$ *and* $\|.\|_2$. *In particular,* $\|Q_1(r)\|_\infty = 1$ *and* $\|Q_1(r)\|_2 < 1$.

Proof. Let us firstly consider the norm $\|.\|_\infty$. If $0 < r \le 1/2$, then the coefficients of $Q_1(r)$ are non-negative, and by theorem 1.5.1, $\|Q_1(r)\|_\infty = 1$, which implies stability in this norm. For the norm $\|.\|_2$, since Q_1 is symmetric, the same theorem implies that $\|Q_1\|_2 = \max_j |\lambda_j(Q_1)|$. We can check that the eigenvectors of Q_1 are given by

$$s_k = \left(\sin\left(\frac{kj\pi}{(N+1)}\right)\right)_{j=1,\dots,N}, \quad k = 1,\dots,N,$$

which are simply the values of the functions $x \mapsto \sin(k\pi x)$ at the points x_j, i.e. the eigenfunctions of the operator $u \mapsto u''$ which vanish at the boundary of Ω. The corresponding eigenvalues λ_k of Q_1 are given by

$$\lambda_k(Q_1) = 1 - 4r\sin^2\left(\frac{k\pi}{2(N+1)}\right), \; 1 \le k \le N.$$

If $0 < r \le 1/2$, we have $|\lambda_k(Q_1)| < 1$ for every k, and the method is stable. □

The condition for stability $0 < r \le 1/2$ is expressed as a constraint on the time-step

$$0 < \Delta t \le \Delta x^2/2\mu. \tag{3.11}$$

Under this condition, the method is thus convergent and

$$\sup_{m;m\Delta t\le T} \|\overline{U}_h^m - U_h^m\|_p \le c(\Delta x^2 + \Delta t),$$

for $p = 2$ and $p = \infty$, where c only depends on u.

From our bound on the consistency error, we see that the method is of order 2 (in space) and 1 (in time). Therefore, the overall error is finally given by $O(\Delta x^2)$ and is not deteriorated by the order 1 in time (we consider $\mu > 0$ fixed, which is a priori not very small compared to Δx). However, this estimate assumes a condition which limits the size of the time step, here (3.11), and we thus say that the method is *conditionally stable*. This is a consequence of the explicit nature of the method.

Let us now move on to the study of our implicit **method 2**. We note here that the exact solution satisfies

$$(1+2r)u(x_j,t_{n+1}) - r(u(x_{j+1},t_{n+1}) + u(x_{j-1},t_{n+1})) = u(x_j,t_n) + \Delta t f_j^{n+1} + \Delta t\overline{\kappa}_j^n,$$

which again defines the consistency error of our method $\overline{\kappa}_j^n$ and we can verify that we still have

$$|\overline{\kappa}_j^n| \leq C(\Delta x^2 + \Delta t),$$

where the constant C depends on $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^4 u}{\partial x^4}|$, $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^2 u}{\partial t^2}|$. If we define the matrix Q by $Q = Q_2^{-1}(r)$, and if we take $K_h^n = Q\overline{K}_h^n$ and $G_h^n = QF_h^{n+1}$, the method can be written in the general form (3.10).

The stability of the method is guaranteed in the following proposition.

Proposition 3.1.4 *The method (3.5) is stable in the norm $||.||_\infty$ and in $||.||_2$. In particular $||Q||_\infty \leq 1$ and $||Q||_2 < 1$.*

Proof. Let us firstly consider the norm $||.||_\infty$. We see that Q_2 is an M-matrix and that the sum of each row is greater than 1. By proposition 2.2.2, we thus have

$$||Q||_\infty = ||Q_2^{-1}||_\infty \leq 1.$$

which implies the stability in this norm. For the norm $||.||_2$, we find that Q_2 admits the same eigenvectors as Q_1, and we can also compute its eigenvalues

$$\lambda_k(Q_2) = 1 + 4r\sin^2(\frac{k\pi}{2(N+1)}),\ 1 \leq k \leq N.$$

Because Q_2 is symmetric, the same holds true for Q and we thus have $||Q||_2 = \max_j |\lambda_k(Q)|$, and the eigenvalues of Q are the reciprocals of $\lambda_k(Q_2)$, i.e. the values

$$\lambda_k(Q) = \left(1 + 4r\sin^2\left(\frac{j\pi}{2(N+1)}\right)\right)^{-1},\ 1 \leq k \leq N.$$

We thus have $|\lambda_k(Q)| < 1$, thus $||Q||_2 \leq 1$ and the method is stable in this norm. □

The method 2 is thus unconditionally stable, and also consistent, therefore convergent.

3.2 Approximation of the transport equation

We are now interested in the transport equation, which is also called advection equation or convection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad x \in]0,1[, \quad t > 0, \tag{E}$$

with various possible boundary conditions depending on the sign of a as we saw in chapter 1.

We denote by Δx the spatial step size and by Δt the time step size, and we let $x_j = j\Delta x$ and $t_n = n\Delta t$ with $0 \leq j \leq N+1$ and $0 \leq n \leq M = T/\Delta t$. We take

$$\lambda = \frac{\Delta t}{\Delta x}.$$

We denote by u_j^n the approximation to $u(x_j, t_n)$ which we would like to compute, and similarly to the heat equation, we denote by U_h^n the vector which contains these values. In the case of periodic boundary conditions we take more specifically $U_h^n = (u_0^n, \ldots, u_N^n)^T$ because the value u_{N+1}^n is given to be equal to u_0^n. Similarly, we take $U_h^n = (u_1^n, \ldots, u_{N+1}^n)^T$ in the case of left sided boundary conditions when $a > 0$ because u_0^n is given, and $U_h^n = (u_0^n, \ldots, u_N^n)^T$ in the case of right sided boundary conditions if $a < 0$.

The methods, for solving our PDE, differ here in the way in which time and spatial derivatives are approximated. If it is natural to make the approximation

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

for the time derivative, there is no reason, a priori, of prioritising one spatial approximation over another:

1. centered approximation:
$$\frac{\partial u}{\partial x}(x_j, t_n) \simeq \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x},$$
2. right-sided approximation:
$$\frac{\partial u}{\partial x}(x_j, t_n) \simeq \frac{u_{j+1}^n - u_j^n}{\Delta x},$$
3. left-sided approximation:
$$\frac{\partial u}{\partial x}(x_j, t_n) \simeq \frac{u_j^n - u_{j-1}^n}{\Delta x}.$$

Not to mention, of course, the multitude of other ways in which we can approximate a first order derivative, for example using more than just two points. Of course, we have to be careful with the boundary conditions which can be periodic or with given values at $j = 0$ or $j = N + 1$, as we explained in the previous section.

Let us start with some examples of explicit schemes for the transport equation.

1. **Centred method**
   Let us begin with a fairly natural method. We will see later on that it is not stable and therefore not used in practice
$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n).$$
   This method is obtained using a centered approximation to the spatial derivative.

2. **Non-centered schemes**

   (a) **Left-sided method**
$$u_j^{n+1} = u_j^n - \lambda a(u_j^n - u_{j-1}^n).$$
   This method is obtained by taking a left-sided approximation to the spatial derivative.

(b) **Right-sided method**

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

This method is obtained by taking a right-sided approximation to the spatial derivative.

3. **The Lax–Friedrichs method**

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

We recognise a modification of the centered method where u_j^n is replaced by the average of u_{j-1}^n and u_{j+1}^n.

4. **The Lax–Wendroff method**

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

We recognise a modification of the centered method where we have added a strange term which contains a discretisation of the second order derivative in space even though there is no second order derivative in the original PDE.

We associate with every method a "stencil". The stencil associated with the point x_j is the set of points which are used to compute u_j^{n+1}, more precisely, it is the set of points x_k, such that $u_j^{n+1} = \sum_k c_k u_k^n$. See figure 3.1.

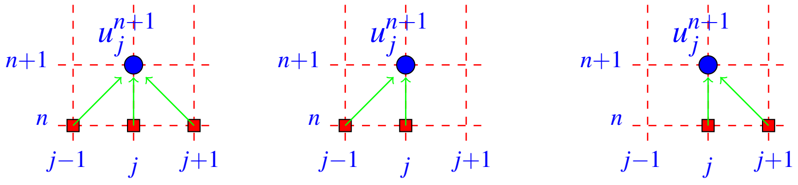

Figure 3.1: The stencils of several methods (from left to right): centered scheme, left-sided scheme and right-sided scheme. The values (at time t_n) marked with squares allow the computation of the value (at time t_{n+1}) marked with a circle.

All of the presented methods are methods using 3 points, i.e. they use, at most, the three points x_{j-1}, x_j and x_{j+1} in order to advance in time. They can be written in the form

$$u_j^{n+1} = H(u_{j-1}^n, u_j^n, u_{j+1}^n), \tag{3.12}$$

where we have introduced a function H called the "discrete solution map". In the linear case, which we are interested in here, the function H is always a linear combination of the values u_k^n. In general, we may study methods with $2L+1$ points. These methods can be written in the general form

$$u_j^{n+1} = H(u_{j-L}^n, \cdots, u_{j+L}^n) \tag{3.13}$$

with the "discrete solution map"

$$H(u^n_{j-L},\cdots,u^n_{j+L}) = \sum_{l=-L}^{+L} c_l u^n_{j+l} \tag{3.14}$$

for some coefficients c_ℓ which depend on λ and a. For $L=1$, we recover the 3-point methods, for $L=2$, we obtain 5-point methods, ...

Here are the coefficients c_ℓ for the formula (3.14) for the 3-point methods mentioned above.

1. Centred method
$$c_{-1} = \frac{\lambda a}{2}, \quad c_0 = 1, \quad c_1 = -\frac{\lambda a}{2}.$$

2. Non-centered methods

   (a) Left-sided method
   $$c_{-1} = \lambda a, \quad c_0 = 1 - \lambda a.$$

   (b) Right-sided method
   $$c_0 = 1 + \lambda a, \quad c_1 = -\lambda a.$$

3. The Lax–Friedrichs method
$$c_{-1} = \frac{1+\lambda a}{2}, \quad c_1 = \frac{1-\lambda a}{2}.$$

4. The Lax–Wendroff method
$$c_{-1} = \frac{\lambda a}{2}(\lambda a + 1), \quad c_0 = 1 - \lambda^2 a^2, \quad c_1 = \frac{\lambda a}{2}(\lambda a - 1).$$

The function H must satisfy the condition

$$H(v,\cdots,v) = v \tag{3.15}$$

which corresponds to the fact that constant states are propagated exactly by the "exact solution map". In other words, if the solution is constant at time t_n, it must remain constant at time t_{n+1}. The relation (3.15) means for a linear method that

$$\sum_{\ell=-L}^{L} c_\ell = 1, \tag{3.16}$$

which we shall assume for the remainder of this discussion.

Note that for a linear method (which is the case for all methods presented here) the relation (3.13) can also be written as

$$U_h^{n+1} = QU_h^n, \tag{3.17}$$

where Q is a matrix, whose constant diagonals are given by the coefficients $c_{-L},\ldots,c_L$ and whose first and last rows must take into account the boundary conditions imposed at $j=0$ and

$j = N+1$.

Let us consider the case of a linear or non-linear equation in conservative form

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}[f(u)] = 0, \quad x \in \mathbb{R}, \quad t > 0 \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{cases}$$

where f is the flux function.

By introducing the midpoints $x_{j+1/2} = \frac{1}{2}(x_j + x_{j+1})$ and the cells $C_i =]x_{i-1/2}, x_{i+1/2}[$ we note that when Δx is small the values $u(x_j, t_n)$ are close to the average values

$$m_j^n := \frac{1}{\Delta x} \int_{C_j} u(x,t_n) dx.$$

According to the conservation law, their values evolve exactly according to the relation

$$m_j^{n+1} - m_j^n = \frac{1}{\Delta x} \int_{t_n}^{t_{n+1}} [f(u(x_{j+1/2},t) - f(u(x_{j-1/2},t)] dt.$$

The right hand side is not computable but can be approximated by

$$\frac{\Delta t}{\Delta x}(f(u(x_{j+1/2},t_n) - f(u(x_{j-1/2},t_n)) \approx \lambda(g(u_j^n, u_{j+1}^n) - g(u_{j-1}^n, u_j^n)),$$

where g is a function of two variables called the numerical flux. This numerical flux must be consistent with the flux f of the PDE in the sense that $g(v,v) = f(v)$.

This leads to what is called the *finite volume method*

$$u_j^{n+1} = u_j^n - \lambda[g(u_j^n, u_{j+1}^n) - g(u_{j-1}^n, u_j^n)], \tag{3.18}$$

which corresponds to a particular choice $H(u,v,w) = v - \lambda[g(v,w) - g(u,v)]$. We note that this choice automatically ensures that $H(v,v,v) = v$ and the conservation of the total mass $\sum_j u_j^{n+1} = \sum_j u_j^n$.

In the linear case ($f(u) = au$), all of the finite difference methods which we have introduced above can be interpreted as finite volume methods with a suitable numerical flux function:

1. Centred method: $g_C(v,w) = a(v+w)/2$.

2. Non-centered method : $g_D(v,w) = av$ in the case $a > 0$ ($= aw$ if $a < 0$).

3. Lax–Friedrichs method: $g_{LF}(v,w) = \frac{a}{2}(w+v) - \frac{1}{2\lambda}(w-v)$.

4. Exercise: determine the numerical flux function associated with the Lax–Wendroff method.

3.3 Analysis of the methods

The numerical analysis of the methods aims to study their convergence, in an appropriate norm, and when they converge, to estimate the error, calculated (often) in the same norm, between the exact solution and the approximated solution. As we have already seen for the heat equation, two important properties come into play: stability and consistency, from which we can then deduce convergence.

Let us begin with the study of consistency. Like for the heat equation, we adopt the following notation.

Definition 3.3.1 *We call the real number κ_j^n given by*

$$\kappa_j^n = \frac{1}{\Delta t}\left(u(x_j,t_{n+1}) - H(u(x_{j-L},t_n),\cdots,u(x_{j+L},t_n))\right)$$

the consistency error of the method (3.13), at the point x_j and time t_n. The consistency error of the method at time t_n is the vector K_h^n whose components are the κ_j^n, and which thus satisfies

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t K_h^n.$$

For a given vector norm $\|\cdot\|$, we say the method is consistent if its consistency error converges to 0 when the step sizes of the discretisation Δt and Δx go to 0. Moreover, we say the method is consistent of order p in space and q in time if and only if there is a constant $C>0$ such that

$$\sup_{n\Delta t \le T} \|K_h^n\| \le C(\Delta x^p + \Delta t^q),$$

for every sufficiently regular solution u.

From the three Taylor series expansions below, we can estimate the consistency errors of the above mentioned schemes in the ℓ^∞-norm, and consequently for the ℓ^2_Δ-norm which is bounded above by the former.

1. For the centered scheme, we have

$$\begin{aligned}\kappa_j^n &= \tfrac{1}{\Delta t}(u(x_j,t_{n+1}) - u(x_j,t_n)) + \tfrac{a}{2\Delta x}(u(x_{j+1},t_n) - u(x_{j-1},t_n)) \\ &= \Big(\partial_t u(x_j,t_n) + O(\Delta t)\Big) + a\Big(\partial_x u(x_j,t_n) + O((\Delta x)^2)\Big).\end{aligned}$$

   Since u is a solution to the PDE, the consistency error is

$$\kappa_j^n = O(\Delta t) + O((\Delta x)^2).$$

   The method is consistent of order 1 in time and order 2 in space.

2. For the left-sided method, we have

$$\kappa_j^n = \frac{1}{\Delta t}(u(x_j,t_{n+1}) - u(x_j,t_n)) + \frac{a}{\Delta x}(u(x_j,t_n) - u(x_{j-1},t_n)) = O(\Delta t) + O(\Delta x).$$

   The method is consistent of order 1 in time and order 1 in space. The same is true for the right-sided method.

3. Lax–Friedrichs method.

   One can show (exercise) that the method is consistent under the condition that $\frac{(\Delta x)^2}{\Delta t}$ tends to 0. In other words, the condition requires that Δx tends to 0 faster than $\sqrt{\Delta t}$.

4. Lax–Wendroff method.

   One can show (exercise) that the consistency error of the Lax–Wendroff method satisfies

$$\kappa_j^n = O((\Delta t)^2) + O((\Delta x)^2).$$

   Thus this is a consistent method of order 2 in time and order 2 in space.

Let us now move on to the study of stability.

Definition 3.3.2 *We say that a method is stable in a vector norm $\|.\|$ if and only if there is a constant C_0 (which may depend on T), such that*

$$\sup_{n\Delta t \leq T} \|Q^n\| \leq C_0.$$

Of course, Lax's equivalence principle (theorem 3.1.1) which we already stated for the heat equation and which told us that "consistency and stability imply convergence" remains valid for the transport equation with the same proof. We will start with the study of stability in the ℓ^∞-norm, presenting (in the case of periodic boundary conditions) some sufficient conditions which permit a very simple proof.

Proposition 3.3.1 *A linear method (3.13)-(3.15) whose coefficients are non-negative is stable in ℓ^∞.*

Proof. Every row in the matrix Q has the coefficients $c_{-L}, \ldots, c_L$, which implies that

$$\|Q\|_\infty = \sum_{l=-L}^{L} |c_l|.$$

If the coefficients are non-negative we thus have

$$\|Q\|_\infty = \sum_{l=-L}^{L} c_l = 1,$$

by the relation (3.16). □

Remark 3.3.1 *A function H which satisfies the assumptions of proposition 3.3.1 is increasing in each of its arguments. Therefore, if two initial conditions V_h^0 and W_h^0 satisfy $V_h^0 \geq W_h^0$ in the sense that $v_j^0 \geq w_j^0$ for every j, then $V_h^n \geq W_h^n$ for every n. This is the discrete analogue of the monotonicity property*

$$v_0 \leq w_0 \implies v(\cdot,t) \leq w(\cdot,t), \quad t > 0,$$

satisfied by the solutions of the transport PDE. We also say that the method is monotone.

Applications of proposition 3.3.1. We recall that $\lambda = \Delta t/\Delta x$.

1. Centred method. We cannot use proposition 3.3.1 because the coefficients of the linear combination cannot all be non-negative. In fact, this method turns out to be numerically unstable.

2. Non-centered methods

   (a) Left-sided method. For $a > 0$ and under the condition that

   $$a\frac{\Delta t}{\Delta x} \leq 1,$$

   the left-sided method is stable in the ℓ^∞-norm.

   (b) Right-sided method. For $a < 0$ and under the condition that $-a\lambda \leq 1$, the right-sided method is stable in the ℓ^∞-norm.

   (c) Note the inconsistency in the choice of the left-sided method if the velocity $a < 0$: the characteristic originating from point (x_j, t_{n+1}) does not intersect the stencil of this method (see figure 3.1). An analogous remark holds true for the right-sided method. We note that the left-sided method with positive velocity and the right-sided method with negative velocity (i.e. taking into account the "direction of the wind") is stable in the ℓ^∞-norm under the condition

   $$|a|\frac{\Delta t}{\Delta x} \leq 1, \tag{3.19}$$

   which is called the CFL condition.[1] The method is called "upwind": it takes into account the direction of the wind. If we do not specify the sign of a, we can write

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

   where $a_- = \min(a, 0)$ and $a_+ = \max(a, 0)$. Noting that $a = a_+ + a_-$ and $|a| = a_+ - a_-$, we can write the upwind method in the form

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

3. Lax–Friedrichs method. Exercise: Show that the Lax–Friedrichs method is stable in the ℓ^∞-norm, under the condition (3.19).

4. Lax–Wendroff method. We cannot use proposition 3.3.1 because the coefficients of the linear combination cannot all be non-negative.

The Lax–Wendroff method is not stable in the ℓ^∞-norm, however it is a "promising" method because it is the only one, among those which we have presented, which is of order 2 in space and time. For this scheme (like some others!), it is necessary to measure the stability in a norm

[1] Initials of three mathematicians which worked on PDEs and/or their numerical solution: Richard Courant, Kurt Friedrichs, and Hans Lewy.

different from $\|.\|_\infty$. The ℓ^2-norm is a good candidate.

For a linear method, the ℓ^2-stability can be characterised with the help of the discrete Fourier transform. We consider here again periodic boundary conditions

$$u(x+1,t) = u(x,t).$$

In particular

$$u(x_{j+N+1},t) = u(x_j,t), \quad j \in \mathbb{Z}.$$

In order to translate this periodicity to the method we take vectors U_h^n whose components are $(u_0^n,\dots,u_N^n)$ and which can be extended by taking the values u_j^n to the whole of $j \in \mathbb{Z}$ using the periodicity condition

$$u_{j+N+1}^n = u_j^n.$$

This allows us to give a precise meaning to the relation

$$u_j^{n+1} = \sum_{l=-L}^{L} c_l u_{j+l}^n, \quad j = 0,\dots,N.$$

We see in particular that the matrix Q such that $U_h^{n+1} = QU_h^n$ is a circulant matrix whose entries are given by

$$q_{i,j} = c_{(j-i)\%(N+1)}.$$

For every vector $V = (v_0,\dots,v_N)^T \in \mathbb{R}^{N+1}$ we can now define the discrete Fourier transform $\hat{V} = (\hat{v}_0,\dots,\hat{v}_N)$ by

$$\hat{v}_k = \frac{1}{\sqrt{N+1}} \sum_{j=0}^{N} v_j \exp\Big(-i2\pi \frac{kj}{N+1}\Big) = \frac{1}{\sqrt{N+1}} \sum_{j=0}^{N} v_j \exp(-i2\pi kj\Delta x) = \langle V, F_k \rangle,$$

where F_k is the vector with entries $\big(\frac{1}{\sqrt{N+1}} \exp(-i2\pi kj\Delta x)\big)_{j=0,\dots N}$. Note that the vector F_k can be naturally extended by periodicity to all of $\mathbb{Z}$, and that the $\hat{v}_k$ are, in general, complex numbers. It is easy to check (exercise) that $(F_0,\dots,F_N)$ is an orthonormal basis of $\mathbb{C}^{N+1}$ with the usual Hilbertian scalar product, and thus

$$V = \sum_{k=0}^{N} \hat{v}_k F_k,$$

which can also be read as the expression for the inverse Fourier transform

$$v_j = \frac{1}{\sqrt{N+1}} \sum_{k=0}^{n} \hat{v}_k \exp(i2\pi kj\Delta x).$$

We have in particular Parseval's identity for the norm $\|\cdot\| = \|\cdot\|_2$:

$$\|V\|^2 = \sum_{j=0}^{N} |v_j|^2 = \sum_{k=0}^{N} |\hat{v}_k|^2 = \|\hat{V}\|^2.$$

The action of Q on a vector V extended by periodicity can be written as

$$(QV)_j = \sum_{l=-L}^{L} c_l v_{j+l},$$

which yields in particular that

$$\begin{aligned}(QF_k)_j &= \tfrac{1}{\sqrt{N+1}} \textstyle\sum_{l=-L}^{L} c_l \exp(i2\pi k(j+l)\Delta x) \\ &= \left(\textstyle\sum_{l=-L}^{L} c_l \exp(i2\pi kl\Delta x)\right) \exp(i2\pi kj\Delta x) \\ &= \alpha_k (F_k)_j,\end{aligned}$$

with

$$\alpha_k := \sum_{l=-L}^{L} c_l \exp(i2\pi kl\Delta x).$$

This conveys the fact that F_k are eigenfunctions of circulant matrices: we have $QF_k = \alpha_k F_k$. In particular, this means that

$$\hat{u}_k^{n+1} = \alpha_k \hat{u}_k^n.$$

We say that α_k is the amplification coefficient of the k^{th} Fourier mode for the method under consideration. Using Parseval's identity, we obtain

$$\|Q\|_2 = \max_{k=0,\ldots,N} |\alpha_k|,$$

and thus that

$$\|Q^m\|_2 = \max_{k=0,\ldots,N} |\alpha_k|^m.$$

This implies that the method is stable in the ℓ^2-norm if and only if

$$|\alpha_k| \leq 1, \quad k = 0, \ldots, N.$$

This condition, which ensures that the Fourier coefficients are bounded by

$$|\hat{u}_k^n| \leq |\hat{u}_k^0|, \quad n \in \mathbb{N},$$

is called *von Neumann stability criterion.*

Example 3 *Let us consider an explicit method as a discretisation of the heat equation as studied in the previous section, assuming periodic boundary conditions. Without any source term, the method takes the form*

$$u_j^{n+1} = (1-2r)u_j^n + ru_{j-1}^n + ru_{j+1}^n, \quad r = \mu\frac{\Delta t}{\Delta x^2}.$$

We thus deduce that

$$\alpha_k = 1 - 2r\Big(1 - \frac{1}{2}(\exp(i2\pi k\Delta x) + \exp(-i2\pi k\Delta x))\Big) = 1 - 2r(1-\cos(2\pi k\Delta x)) = 1 - 4r\sin^2(\pi k\Delta x)).$$

We can see that the von Neumann stability criterion is satisfied if and only if $0 < r \leq 1/2$, *which is the same result as we had obtained in the case of homogeneous Dirichlet boundary conditions. For the implicit scheme (still for the heat equation with periodic boundary conditions), we can show that*

$$\alpha_k = \frac{1}{1+4r\sin^2(k\pi\Delta x)}.$$

This method is thus unconditionally stable in the ℓ^2*-norm, i.e. without any conditions on the discretisation mesh sizes* Δt *and* Δx*. We have already seen a similar result in the case of homogeneous Dirichlet boundary conditions.*

Let us now return to the transport equation and in particular to methods (3.13), (3.14), (3.16).

1. For the centered method, we have

   $$\alpha_k = 1 + \frac{\lambda a}{2}(e^{-i\theta} - e^{i\theta}) = 1 - i\lambda a \sin\theta,$$

   where we denoted by $\theta = 2k\pi\Delta x$. We thus deduce that $|\alpha_k| > 1$, for any k such that $\sin\theta \neq 0$. The method is therefore not ℓ^2-stable.

2. For the upwind method (left-sided if $a > 0$ and right-sided if $a < 0$), we find in the case $a > 0$

   $$\alpha_k = 1 - \lambda a + \lambda a \cos\theta - i\lambda a \sin\theta.$$

   Exercise: Show that the upwind method is ℓ^2-stable under the following CFL condition

   $$|a|\frac{\Delta t}{\Delta x} \leq 1. \tag{3.20}$$

3. For the Lax–Friedrichs method, we find

   $$\alpha_k = \cos\theta - i\lambda a \sin\theta.$$

   Exercise: Show that the Lax–Friedrichs method is ℓ^2-stable under the same CFL condition.

4. For the Lax–Wendroff method, we find

   $$\alpha_k = 1 - 2(\lambda a)^2 \sin^2\theta - i\lambda a \sin(2\theta).$$

   where we denoted $\theta = k\pi\Delta x$. Exercise: Show that the Lax–Wendroff method is ℓ^2-stable under the same CFL condition.

Remark 3.3.2 (Concerning the CFL condition (3.20)) *The quantity* $|a|\Delta t$ *is the distance travelled during time* Δt *at velocity a. Since the exact solution satisfies* $u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$*, the CFL condition expresses the fact that the characteristic curve starting at the point* x_j *at time* t_{n+1} *must intersect with the line* $t = t_n$ *at a point contained in the stencil of the method. See figure 3.2.*

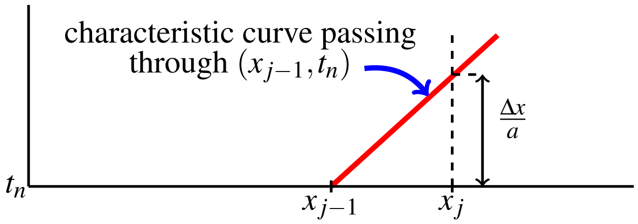

Figure 3.2: Interpretation of the CFL condition (3.20). Case $a > 0$. If Δt is smaller than $(\Delta x)/a$, then the characteristic passing through (x_j, t_{n+1}) intersects with the line $t = t_n$ in a point located between x_{j-1} and x_j.

We conclude by returning to the notion of the order of approximation of the methods. We assume in the following that the ratio

$$\lambda = \frac{\Delta t}{\Delta x}$$

is constant and thus the time step size and spatial step size tend to 0 at the same speed.

Definition 3.3.3 *We say that a method is of order p, if for every sufficiently regular solution of the equation (1.4), we have*

$$\kappa_j^n = O((\Delta t)^p) = O((\Delta t)^p + (\Delta x)^p). \tag{3.21}$$

Let us write out the Taylor series of the solution, assuming sufficient regularity. We use the following notation here: $\overline{u}_j^n = u(x_j, t_n)$, $\partial_x \overline{u}_j^n = \partial_x u(x_j, t_n)$, etc. We have on the one hand side that

$$H(\overline{u}_{j-L}^n, \cdots, \overline{u}_{j+L}^n) = \left[\sum_{\ell=-L}^{L} c_\ell\right] \overline{u}_j^n + \left[\sum_{\ell=-L}^{L} \ell \Delta x c_\ell\right] \partial_x \overline{u}_j^n + \left[\sum_{\ell=-L}^{L} \frac{(\ell \Delta x)^2}{2} c_\ell\right] \partial_{xx} \overline{u}_j^n + \cdots$$

On the other hand, by the relations $u_t = -a u_x$ and $u_{tt} = a^2 u_{xx}$, we have

$$\overline{u}_j^{n+1} = \overline{u}_j^n - a(\Delta t)\partial_x \overline{u}_j^n + \frac{(a\Delta t)^2}{2} \partial_{xx} \overline{u}_j^n + \cdots$$

Hence the Taylor series expansion of the consistency error is (taking note of (3.16))

$$\kappa_j^n = -\left[a + \frac{1}{\lambda} \sum_{\ell=-L}^{L} \ell c_\ell\right] \partial_x \overline{u}_j^n + \left[\frac{a^2 \Delta t}{2} - \frac{1}{\lambda} \sum_{\ell=-L}^{L} \frac{\ell^2 \Delta x}{2} c_\ell\right] \partial_{xx} \overline{u}_j^n + \cdots$$

For a constant $\lambda = \frac{\Delta t}{\Delta x}$ we can thus write

$$\kappa_j^n = -\left[a + \frac{1}{\lambda} \sum_{\ell=-L}^{L} \ell c_\ell\right] \partial_x \overline{u}_j^n + \frac{\Delta t}{2}\left[a^2 - \frac{1}{\lambda^2} \sum_{\ell=-L}^{L} \ell^2 c_\ell\right] \partial_{xx} \overline{u}_j^n + O((\Delta t)^2).$$

Hence we deduce the following result

Proposition 3.3.2 *If the coefficients c_ℓ of a method (3.13), (3.14), (3.16) satisfy $\sum_{\ell=-L}^{L} \ell c_\ell = -\lambda a$ then the method is of order at least equal to* 1. *If, in addition, $\sum_{\ell=-L}^{L} \ell^2 c_\ell = (\lambda a)^2$, then it is of order at least equal to* 2.

Proposition 3.3.3 *A linear conservative 3-point method (here conservative means that it can be written in the form (3.18)), associated to a consistent numerical flux is of order at least equal to 1.*

Proof. Let us show that $\sum_{\ell=-L}^{L} \ell c_\ell = -\lambda a$, i.e. for a 3-point method that is $c_1 - c_{-1} = -\lambda a$. If $g(v,w) = \alpha v + \beta w$ is the numerical flux associated with the method, its coefficients are thus $c_{-1} = \lambda\alpha$, $c_0 = 1 - \lambda(\alpha - \beta)$ and $c_1 = -\lambda\beta$. The desired relation then follows from the consistency of the flux g. □

Proposition 3.3.4 *There is only one conservative 3-point method, whose associated numerical flux is consistent and which is of order two. This method is given by the Lax–Wendroff method.*

Proof. If the numerical flux is $g(v,w) = \alpha v + \beta w$, then the method can be written as

$$u_j^{n+1} = c_{-1} u_{j-1}^n + c_0 u_j^n + c_1 u_{j+1}^n$$

with $c_{-1} = \alpha\lambda, c_0 = 1 - \alpha\lambda + \beta\lambda$ and $c_1 = -\beta\lambda$. We have $c_{-1} + c_0 + c_1 = 1$ and as the flux is consistent, $c_{-1} - c_1 = \lambda a$. Such a method can therefore be written in terms of a single parameter which we take to be $q = c_{-1} + c_1 = 1 - c_0$. The method can thus be expressed as follows

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{q}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \tag{3.22}$$

By proposition 3.3.2, this method is of order two if $c_{-1} + c_1 = (\lambda a)^2$, i.e. if $q = \lambda^2 a^2$. We recognise immediately the Lax–Wendroff method. We can check that this method is not of order higher than two. □

Let us consider methods of the form (3.22) where the real number q is a parameter which we call the *viscosity coefficient* of the method (why?). The method obtained for $q = (\lambda a)^2$ is thus the Lax–Wendroff method.

Proposition 3.3.5 *A linear conservative 3-point method (3.13) whose numerical flux is consistent, is stable in ℓ^2 if and only if its viscosity coefficient q satisfies*

$$(\lambda a)^2 \le q \le 1. \tag{3.23}$$

Proof. The amplification coefficient of the method (3.22) is given by (we take $c = \lambda a$)

$$\alpha_k = 1 - \frac{c}{2}(\exp(i2\pi k\Delta x) - \exp(-i2\pi k\Delta x)) + \frac{q}{2}(\exp(i2\pi k\Delta x) - 2 + \exp(-i2\pi k\Delta x))$$

$$= 1 - q + q\cos(2\pi k\Delta x) - ic\sin 2\pi k\Delta x$$

whence

$$\alpha_k = 1 - q(1 - \cos(2\pi k\Delta x)) - 2ic\cos(2\pi k\Delta x/2)\sin(2\pi k\Delta x/2).$$

Taking $y = (\sin(2\pi k\Delta x/2))^2$, we have

$$|\alpha_k|^2 = (1 - 2qy)^2 + 4c^2y(1-y).$$

The von Neumann criterion is thus satisfied if and only if

$$0 \le (1-2qy)^2 + 4c^2y(1-y) \le 1, \qquad \forall y \in [0,1].$$

These inequalities are satisfied if

$$-qy + q^2y^2 + c^2y(1-y) \le 0, \qquad \forall y \in [0,1].$$

Given that $y \in]0,1]$, the inequalities are satisfied if

$$-q + q^2y + c^2(1-y) \le 0.$$

We recognise an affine function of y, which is thus non-positive on an interval if and only if it is non-positive on both endpoints of this interval. Hence the result follows. □