6.7 The explicit Euler three point finite difference scheme for the heat equation

We now turn to numerical approximation methods, more specifically finite difference methods. We concentrate on the 1d problem (6.1) with $g = 0$, *i.e.*, homogeneous Dirichlet boundary conditions. We assume that the solution u is as regular as we need it to be.

The general idea of finite difference methods is to replace derivatives by difference quotients involving approximate discrete unknowns, and solve for these unknowns. In the case of an evolution equation, we need a space-time grid. Let us thus be given two positive integers N and M. We note $h = \Delta x = \frac{1}{N+1}$ and $x_i = ih$ for $i = 0, 1, \ldots, N+1$. Similarly, we note $k = \Delta t = \frac{T}{M+1}$ and $t_j = jk$ for $j = 0, 1, \ldots, M+1$. The parameter h is called the space grid step and the parameter k the time grid step, or time step. The grid points are the points (x_i, t_j). Eventually, we will let N and M go to infinity, or equivalently, h and k go to 0.

The discrete unknowns are scalars u_i^j for the above values of i and j, and it is hoped that u_i^j will be an approximation of $u(x_i, t_j)$, that should become better and better as N and M are increased (in a perfect world). The boundary condition can be enforced exactly by requiring that

$$u_0^j = u_{N+1}^j = 0$$

for all $j = 0, \ldots, M+1$. The initial condition is naturally discretized by requiring that

$$u_i^0 = u_0(x_i)$$

for all $i = 0, \ldots, N+1$. Note that boundary values and initial data are consistent with each other in the sense that $u_0^0 = u_0(0) = u_0(1) = u_{N+1}^0$.

The right-hand side of the equation is discretized by setting $f_i^j = f(x_i, t_j)$.

The only values that are left unknown at this stage are thus u_i^j for $i = 1, \ldots, N$ and $j = 1, \ldots, M+1$. We also use the notation

$$U^j = \begin{pmatrix} u_1^j \\ u_2^j \\ \vdots \\ u_N^j \end{pmatrix} \in \mathbb{R}^N$$

to denote the vector of approximate values on the space grid at time t_j. Note a fundamental difference with variational approximation methods such as the finite element method, which is that the approximation is not a function, but a finite set of values.

In the explicit Euler three point scheme, the derivatives are approximated as follows. For the time derivative of the exact solution, we have the forward differential quotient approximation

$$\frac{\partial u}{\partial t}(x_i,t_j) \approx \frac{u(x_i,t_{j+1})-u(x_i,t_j)}{k},$$

and for the second space derivative, by combining a forward and a backward differential quotient, we obtain the central approximation

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}(x_i,t_j) &\approx \frac{\frac{u(x_{i+1},t_j)-u(x_i,t_j)}{h} - \frac{u(x_i,t_j)-u(x_{i-1},t_j)}{h}}{h} \\ &= \frac{u(x_{i-1},t_j)-2u(x_i,t_j)+u(x_{i+1},t_j)}{h^2},\end{aligned}$$

which can be given a precise meaning by using Taylor expansions (we will come back to that later). The finite difference method mimics these approximations by replacing the exact values of the solution at the grid points by the discrete unknowns. In this particular case, we end up with the following scheme:

$$\begin{cases} \dfrac{u_i^{j+1}-u_i^j}{k} - \dfrac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h^2} = f_i^j \text{ for } i=1,\dots,N, j=0,\dots,M, \\ u_i^0 = u_0(x_i) \text{ for } i=1,\dots,N, \\ u_0^j = u_{N+1}^j = 0 \text{ for } j=0,\dots,M+1. \end{cases} \tag{6.15}$$

Of course, at this point, there is no indication that (6.15) has anything to do with (6.1). The name explicit or forward Euler comes from the fact that the time derivative is approximated in the same way as it is in the case of the forward Euler method for ODE's, whereas the three point name comes from the three point centered approximation of the second order space derivative.

We may rewrite the first equation of the scheme in vector form as

$$\frac{U^{j+1}-U^j}{k} + A_h U^j = F^j \text{ for } j=0,\dots,M,$$

where A_h is the $N\times N$ tridiagonal matrix

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix},$$

and F^j the vector

$$F^j = \begin{pmatrix} f_1^j \\ f_2^j \\ \vdots \\ f_N^j \end{pmatrix} \in \mathbb{R}^N.$$

The initial condition is also a vector

$$U_0 = \begin{pmatrix} u_0(x_1) \\ u_0(x_2) \\ \vdots \\ u_0(x_N) \end{pmatrix} \in \mathbb{R}^N.$$

With this notation, the numerical scheme is equivalent to

$$\begin{cases} U^{j+1} = U^j - kA_hU^j + kF^j \text{ for } j = 0, \dots, M, \\ U^0 = U_0. \end{cases} \tag{6.16}$$

This simple recursion formula shows that the scheme is well-defined. We can also note the appearance of the factor $\frac{k}{h^2} = \frac{\Delta t}{\Delta x^2}$ which will play an important role in the sequel.

Let us now introduce a few notions of interest.

Definition 6.7.1 *Let v be a function defined on $[0,1]$. We define the space grid sampling operator S_h by*

$$S_h(v) = \begin{pmatrix} v(x_1) \\ v(x_2) \\ \vdots \\ v(x_N) \end{pmatrix} \in \mathbb{R}^N.$$

Let now u be a solution of problem (6.1). *We define the* truncation error *of the present finite difference method to be the sequence of vectors*

$$\varepsilon(u)^j = \frac{S_h(u_{t_{j+1}}) - S_h(u_{t_j})}{k} + A_hS_h(u_{t_j}) - F^j,$$

where the notation u_t stands for the function $x \mapsto u(x,t)$.

To obtain the truncation error, we just take the finite difference scheme and replace the discrete unknowns with the corresponding grid samplings of a solution of the heat equation. Its name stems from the fact that, if we were to fictitiously apply

the numerical scheme with one time step starting from the exact sampling values at t_j, then we would make an error $S_h(u_{t_{j+1}}) - \widetilde{U}^{j+1} = k\varepsilon(u)^j$. The truncation error is not however directly related to the actual solution error between the sampling of the solution and the discrete unknown, as we will see later.

In order to analyze the convergence of finite difference methods, we need to introduce the function space

$$C^{m,n}(\bar{Q}) = \{u; \forall t \in [0,T], u_t \in C^m([0,1]) \text{ and } \forall x \in [0,1], u_x \in C^n([0,T]) \\ \text{with all derivatives uniformly bounded on } \bar{Q}\}.$$

We equip the space $\mathbb{R}^N$ with the infinity norm

$$\|U\|_{\infty,h} = \max_{1\le i\le N} |U_i|.$$

We have the following easy estimate concerning the truncation error for the explicit Euler three point numerical scheme.

Proposition 6.7.1 *Assume that $u \in C^{4,2}(\bar{Q})$. Then we have*

$$\max_{0\le j\le M} \|\varepsilon(u)^j\|_{\infty,h} \le C(h^2 + k), \tag{6.17}$$

where the constant C depends only on u.

We say that the forward Euler three point scheme is consistent, of order 1 in time and order 2 in space, all these terms to be made precise later on.

Proof. We use Taylor-Lagrange expansions. First we use the fact that u_x is of class C^2. Therefore, for all i and j, there exists $\theta_{ij} \in]t_j, t_{j+1}[$ such that

$$u(x_i,t_{j+1}) = u(x_i,t_j) + k\frac{\partial u}{\partial t}(x_i,t_j) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i,\theta_{ij}),$$

which we rewrite as

$$\frac{u(x_i,t_{j+1}) - u(x_i,t_j)}{k} = \frac{\partial u}{\partial t}(x_i,t_j) + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,\theta_{ij}).$$

Similarly, u_t is of class C^4. Therefore, for all i and j, there exists $\xi_{ij}^+ \in]x_i, x_{i+1}[$, $\xi_{ij}^- \in]x_{i-1}, x_i[$ such that

$$u(x_{i+1},t_j) = u(x_i,t_j) + h\frac{\partial u}{\partial x}(x_i,t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_j) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_j) + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^+,t_j),$$

and

$$u(x_{i-1},t_j) = u(x_i,t_j) - h\frac{\partial u}{\partial x}(x_i,t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_j) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_j) + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^-,t_j),$$

which we rewrite as

$$\frac{u(x_{i-1},t_j)-2u(x_i,t_j)+u(x_{i+1},t_j)}{h^2}=\frac{\partial^2 u}{\partial x^2}(x_i,t_j)+\frac{h^2}{24}\Big(\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^-,t_j)+\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^+,t_j)\Big).$$

Taking into account the boundary conditions $u(x_0,t_j)=u(x_{N+1},t_j)=0$, we see that

$$\varepsilon(u)^j=S_h\Big(\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}\Big)-F^j+R^j$$

with

$$R_i^j=\frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,\theta_{ij})-\frac{h^2}{24}\Big(\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^-,t_j)+\frac{\partial^4 u}{\partial x^4}(\xi_{ij}^+,t_j)\Big).$$

Now we recall that u is a regular solution of the heat equation, therefore due to the definition of the sampling operator, $S_h\big(\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}\big)-F^j=0$, even for $j=0$ by continuity. Moreover

$$|R_i^j|\le\max\Big(\frac{1}{2}\max_{\bar{Q}}\Big|\frac{\partial^2 u}{\partial t^2}\Big|,\frac{1}{12}\max_{\bar{Q}}\Big|\frac{\partial^4 u}{\partial x^4}\Big|\Big)(k+h^2),$$

for all i and j, which concludes the proof of the proposition. □

6.8 The implicit Euler and leapfrog schemes

Before describing and analyzing general finite difference schemes, we give two more simple examples. The first example is the implicit or backward Euler three point scheme, which is associated with the backward differential quotient approximation of the time derivative

$$\frac{\partial u}{\partial t}(x_i,t_j)\approx\frac{u(x_i,t_j)-u(x_i,t_{j-1})}{k},$$

also used under the same name in the context of the numerical approximation of ODE's. In vector form, this scheme reads

$$\frac{U^j-U^{j-1}}{k}+A_hU^j=F^j \text{ for } j=1,\dots,M+1,$$

or equivalently

$$\frac{U^{j+1}-U^j}{k}+A_hU^{j+1}=F^{j+1} \text{ for } j=0,\dots,M.$$

This scheme is called implicit, because the above formula is not a simple recursion formula, but U^{j+1} appears as the solution of an equation once U^j is known. It is not a priori clear that this equation is solvable. In this particular case, we have

$$\begin{cases} U^{j+1} = (I + kA_h)^{-1}(U^j + kF^{j+1}) \text{ for } j = 0, \dots, M, \\ U^0 = U_0, \end{cases} \tag{6.18}$$

since it is not hard to see that the matrix $I + kA_h$ is symmetric, positive definite, hence invertible. In practical terms, the implementation of the backward Euler methods entails the solution of a linear system at each time step, whereas the explicit method is simply a matrix-vector product and vector addition. The implicit method is thus more computationally intensive than the explicit method, but it has other benefits as we will see later.

The analysis of the truncation error of the implicit Euler scheme is basically the same as in the explicit case. The method is likewise consistent, of order 1 in time and order 2 in space.

The second example is the leapfrog or Richardson method, which is associated with the central differential quotient approximation of the time derivative

$$\frac{\partial u}{\partial t}(x_i, t_j) \approx \frac{u(x_i, t_{j+1}) - u(x_i, t_{j-1})}{2k}.$$

In vector form, this scheme reads

$$\frac{U^{j+1} - U^{j-1}}{2k} + A_h U^j = F^j \text{ for } j = 1, \dots, M.$$

This scheme is an explicit two-step method since U^{j+1} is explicitly given in terms of U^j and U^{j-1}.

$$\begin{cases} U^{j+1} = U^{j-1} - 2kA_h U^j + 2kF^j \text{ for } j = 1, \dots, M, \\ U^0 = U_0, U^1 = U_1. \end{cases} \tag{6.19}$$

Of course, since this is a two-step method, a given value U_1 must somehow be ascribed to U^1 in order to initialize the recursion.

The idea behind the leapfrog scheme is that the truncation error is of order 2 in time and order 2 in space, *i.e.*, the truncation error is bounded from above by a quantity of the form $C(h^2 + k^2)$, which would seem to be advantageous as compared to both Euler schemes. Unfortunately, we will see that the improved truncation error is accompanied by instability, which prevents the method from being convergent. It is not usable in practice, and this example shows that a naive approach to finite difference schemes may very well badly fail.

6.9 General finite difference schemes, consistence, stability, convergence

In this section, we introduce a general framework for dealing with finite difference schemes. A finite difference scheme for the heat equation, or for any other linear evolution partial differential equation, is constructed by forming linear combinations of partial differential quotients and replicating these linear combinations on the purely discrete level. It can be cast in the following form: Let us be given two positive integers l and m with $l+m \geq 1$, and a set of $l+m+1$ matrices B_k, $-m \leq k \leq l$, each of size $N \times N$ the entries of which are functions of h and k. We assume that B_l is invertible.

A general $l+m$ step finite difference scheme is then a recursion formula for a sequence of vectors $U^j \in \mathbb{R}^N$, of the form

$$B_l U^{j+l} + B_{l-1} U^{j+l-1} + \cdots + B_0 U^j + \cdots + B_{-m} U^{j-m} = G^j, j \geq m \qquad (6.20)$$

with given initial data

$$U^0 = U_0, U^1 = U_1, \ldots, U^{l+m-1} = U_{l+m-1}.$$

The right-hand side vector G^j is to be constructed from f, but is not necessarily F^j. As before, the intended meaning of U^j is that u_i^j is expected to provide an approximation of $u(x_i, t_j)$.

Definition 6.9.1 *We say that the scheme* (6.20) *is* explicit *is the leading matrix B_l is of the form cI with $c \neq 0$. Otherwise, the scheme is called* implicit.

In an explicit method, the next vector U^{j+l} is thus directly obtained from previously computed vectors by matrix-vector multiplications and vector additions, whereas an implicit method requires to solve a linear system at each time step.

Let us see how the previously introduced schemes fit into this general picture. For forward Euler, we have

$$\begin{cases} \frac{1}{k} U^{j+1} + \left(-\frac{1}{k} I + A_h \right) U^j = F^j, \\ U^0 = U_0, \end{cases}$$

so that $l = 1$, $m = 0$, $B_1 = \frac{1}{k} I$, $B_0 = -\frac{1}{k} I + A_h$ and $G^j = F^j$. It is obviously one step and explicit. Of course, we can also write it with for example $B_1 = I$, $B_0 = -I + kA_h$ and $G^j = kF^j$, there is no uniqueness of the general form for a given scheme. The backward Euler method is

$$\begin{cases} (I + kA_h) U^j - U^{j-1} = kF^j, \\ U^0 = U_0, \end{cases}$$

so that $l = 0$, $m = 1$, $B_0 = I + kA_h$, $B_{-1} = -I$ and $G^j = kF^j$. It is obviously one step and implicit (recall that $I + kA_h$ is invertible). Finally, the leapfrog scheme is

$$\begin{cases} U^{j+1} + 2kA_hU^j - U^{j-1} = 2kF^j, \\ U^0 = U_0, U^1 = U_1. \end{cases}$$

so that $l = 1$, $m = 1$, $B_1 = I$, $B_0 = 2kA_h$, $B_{-1} = -I$ and $G^j = 2kF^j$. It is obviously two step and explicit.

Remark 6.9.1 As mentioned before, there is no uniqueness of a general form for a given scheme. Indeed, given a general form, we can obtain another one by multiplying everything by an arbitrary function of h and k, or even by an arbitrary $N \times N$ invertible matrix function of h and k. So the definition of explicit or implicit scheme as stated before is attached to a general form and not to the scheme under consideration. However, it should be quite clear that writing the backward Euler scheme as

$$U^j - (I + kA_h)^{-1}U^{j-1} = k(I + kA_h)^{-1}F^j$$

and thus declaring it explicit, is somehow cheating. Indeed, the matrix $(I + kA_h)^{-1}$ is not know explicitly[4]. Thus the real issue is an implementation issue: do we need to solve a linear system to compute the scheme, or not? In the former case, the scheme is implicit and the latter case, it is explicit. □

There is nothing in the definition of a general finite difference scheme given above that even alludes to a particular partial differential equation that we might be interested in approximating. We therefore need a way of comparing the vectors $U^j \in \mathbb{R}^N$ and the function u solution of problem (6.1). An obvious idea is to use the sampling operator already introduced in Definition 6.7.1.

Even then, quantitatively comparing two vectors of $\mathbb{R}^N$ involves the choice of a norm on $\mathbb{R}^N$. We are ultimately interested in letting $N \to +\infty$, thus we need a norm for each value of N. There is no reason at this point to do anything else that to choose an arbitrary norm $\|\cdot\|_N$ on $\mathbb{R}^N$ for each N. Two popular choices are

$$\|U\|_{\infty,h} = \max_{1 \le i \le N} |U_i| \text{ and } \|U\|_{2,h} = \sqrt{h}\Big(\sum_{i=1}^{N} U_i^2\Big)^{1/2},$$

(recall that $h = \frac{1}{N+1}$). The reason for the $\sqrt{h}$ factor is for comparison with the L^2 norm in the limit $h \to 0$. Of course, it is well-known that any two norms on $\mathbb{R}^N$ are equivalent, but the constants in the norm equivalence depend on N. For instance,

$$\|U\|_{2,h} \le \|U\|_{\infty,h} \le \frac{1}{\sqrt{h}}\|U\|_{2,h}$$

[4]Well, actually it may well be known somewhere in the literature, but let us assume it is not known for the sake of the argument.

with basically optimal constants.

We can now give a few definitions.

Definition 6.9.2 *Let u be a sufficiently regular solution of problem* (6.1)*. The* truncation error *of the general finite difference method* (6.20) *is the sequence of vectors*

$$\varepsilon(u)^j = B_l S_h(u_{t_{j+l}}) + \cdots + B_0 S_h(u_{t_j}) + \cdots + B_{-m} S_h(u_{t_{j-m}}) - G^j, \tag{6.21}$$

for $j \geq m$.

Again, we just replace the discrete unknown with the grid sampling of the solution in the finite difference scheme formula. Note that, since for any given scheme, there are infinitely many different general formulas describing the same scheme, the truncation error of a given scheme depends on how it is written in general form. Fortunately, this is totally irrelevant for the ensuing analysis. We just need to be careful in the application of the general results in each particular case.

Definition 6.9.3 *We say that the scheme* (6.20) *is* consistent for the family of norms $\|\cdot\|_N$ *if*

$$\max_{j \leq T/k} \|\varepsilon(u)^j\|_N \to 0 \text{ when } (h,k) \to (0,0). \tag{6.22}$$

We say that it is of order p in space and q in time for the family of norms $\|\cdot\|_N$ if

$$\max_{j \leq T/k} \|\varepsilon(u)^j\|_N \leq C(h^p + k^q), \tag{6.23}$$

where the constant C only depends on u.

Consistency means that the scheme is trying its best to locally approximate the right partial differential equation problem in the N norm. Of course, the above definitions depend on the choice of norm and on the choice of general form. A given scheme may well be consistent for one family of norms and not for another, or be of some order in one general form and of another order in another general form. It is up to us to choose the best norm/general form combination. As in the particular cases that we have already seen, checking consistency and computing time and space orders is just a matter of having enough patience to write down the relevant Taylor-Lagrange expansions.

A significantly subtler notion is that of *stability*.

Definition 6.9.4 *Let $\mathscr{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ be such that $(0,0) \in \bar{\mathscr{S}}$. We say that the scheme* (6.20) *is* stable for the family of norms $\|\cdot\|_N$ under condition $\mathscr{S}$*, if there*

exists two constants C_1 and C_2 which only depend on T such that, for all $(h,k) \in \mathscr{S}$ and all initial data U_j,

$$\max_{j \leq T/k} \|U^j\|_N \leq C_1 \max_{0 \leq j \leq l+m-1} \|U_j\|_N + C_2 \max_{j \leq T/k} \|G^j\|_N. \tag{6.24}$$

If $\mathscr{S} = \mathbb{R}_+^* \times \mathbb{R}_+^*$, we say that the scheme is unconditionally stable.

Stability makes no reference to the partial differential equation. It is just a property of the recursion formula which controls the growth of its solutions in terms of the initial data and right-hand side.

Definition 6.9.5 *We say that the scheme* (6.20) *is* convergent for the family of norms $\|\cdot\|_N$ *if*

$$\max_{j \leq T/k} \|U^j - S_h(u_{t_j})\|_N \to 0 \text{ when } (h,k) \to (0,0), (h,k) \in \mathscr{S}. \tag{6.25}$$

If we recall that $S_h(u_{t_j})$ is just a notation for $u(x_i,t_j)$, $i = 1,\ldots,N$, we see that convergence of the scheme means that $|u_i^j - u(x_i,t_j)|$ tends to 0 (at least if the choice of norms $\|\cdot\|_N$ is reasonable enough), or that the computed discrete unknowns u_i^j are in effect approximations of the value of the solution at gridpoints.

The relevance of the above definitions is clarified by means of the Lax theorem:

Theorem 6.9.1 *Assume that*

$$\max_{0 \leq j \leq l+m-1} \|U_j - S_h(u_{t_j})\|_N \to 0 \text{ when } (h,k) \to (0,0), (h,k) \in \mathscr{S}.$$

If the scheme (6.20) *is consistent and stable under condition $\mathscr{S}$ for the family of norms $\|\cdot\|_N$, then it is convergent for that same family of norms.*

Proof. Let us compare the formulas for the truncation error and for the scheme.

$$B_l S_h(u_{t_{j+l}}) + \cdots + B_0 S_h(u_{t_j}) + \cdots + B_{-m} S_h(u_{t_{j-m}}) = \varepsilon(u)^j + G^j,$$
$$B_l U^{j+l} + \cdots + B_0 U^j + \cdots + B_{-m} U^{j-m} = G^j.$$

Setting $V^j = S_h(u_{t_j}) - U^j$ and subtracting the above two formulas, we see that

$$B_l V^{j+l} + \cdots + B_0 V^j + \cdots + B_{-m} V^{j-m} = \varepsilon(u)^j,$$

with the initial data

$$V^j = U_j - S_h(u_{t_j}) \text{ for } 0 \leq j \leq l+m-1.$$

By the stability hypothesis, it follows that

$$\max_{j \leq T/k} \|V^j\|_N \leq C_1 \max_{0 \leq j \leq l+m-1} \|U_j - S_h(u_{t_j})\|_N + C_2 \max_{j \leq T/k} \|\varepsilon(u)^j\|_N,$$

and the right-hand side goes to 0 by consistency and the hypothesis on the initial data for the scheme. □

Remark 6.9.2 We have written here the useful part of the Lax theorem, *i.e.*, consistency plus stability imply convergence. There is a less useful converse part, which says that convergence for all right-hand sides and initial data implies stability and consistency. Therefore, we have not missed anything by focusing on consistency and stability. □

Corollary 6.9.2 *Assume that the scheme is stable under condition $\mathscr{S}$ and of order p in space and q in time for the family of norms $\|\cdot\|_N$, and that*

$$\max_{0\leq j\leq l+m-1} \|U_j - S_h(u_{t_j})\|_N \leq C(h^p + k^q),$$

for some constant C. Then

$$\max_{j\leq T/k} \|U^j - S_h(u_{t_j})\|_N \leq C'(h^p + k^q),$$

where C' only depends on T and u.

Remark 6.9.3 High order stable schemes thus result in (in principle) more accurate approximations than low order schemes. This is conditional on the initial data for the scheme not destroying this accuracy. If the scheme uses several time steps, the corresponding initial data must therefore be computed by using some other, equally accurate method. If the scheme is one time step, then we are liberty to have exact initial data (discounting round-off errors). □

Remark 6.9.4 Let us emphasize again that all this is highly dependent on the choice of norms. Assume that, for one outrageous reason or another, we had chosen $\|U\|_N = 2^{-N}\|U\|_{\infty,h}$. Then, it is likely that even the most wildly non consistent scheme for the ∞,h norms would become consistent for the new norms! Since stability is not affected by multiplication of the norm by a constant, if the scheme was stable for the ∞,h norms, then it would also be stable for the $\|\cdot\|_N$ norms.[5] Hence, by the Lax theorem, it would be convergent for that family of norms. There is however no contradiction. Indeed, saying that $2^{-N}|u_i^j - u(x_i,t_j)| \to 0$ tells us next to nothing about what we really are interested in, namely, is u_i^j a good approximation of $u(x_i,t_j)$ or not. In this respect, the two choices $\|\cdot\|_{\infty,h}$ and $\|\cdot\|_{2,h}$ are much more natural. □

Remark 6.9.5 It should also be noted that the fact that the underlying partial differential equation is the heat equation plays no role in the Lax theorem. The theorem holds true for any finite difference scheme devised to approximate the solution of any evolution partial differential equation problem. □

[5] Note that stability is also affected by the general form used to write the scheme, via the term G^j.

Let us apply the previous results to the explicit Euler scheme. We first recall a few facts about *operator matrix norms*. Let $\|\cdot\|_N$ be a norm on $\mathbb{R}^N$. For any $N\times N$ matrix A, the *induced matrix norm* or *operator norm* is defined by

$$|||A|||_N = \max_{U\in\mathbb{R}^N\setminus\{0\}} \frac{\|AU\|_N}{\|U\|_N}.$$

Of course, we then have

$$\|AU\|_N \leq |||A|||_N \|U\|_N,$$

for all $U\in\mathbb{R}^N$ and the operator norm is the smallest such multiplicative constant on the right. It is well-known that

$$|||A|||_{\infty,h} = \max_{1\leq i\leq N}\Big(\sum_{j=1}^{N}|A_{ij}|\Big)$$

and that

$$|||A|||_{2,h} = \sqrt{\rho(A^TA)},$$

where $\rho(B)$ denotes the *spectral radius* of B, *i.e.*, $\rho(B)=\max\{|\lambda_i|\}$ where $\lambda_i\in\mathbb{C}$ are the eigenvalues of B.

Proposition 6.9.1 *Let $\mathscr{S}=\big\{(h,k)\in\mathbb{R}_+^*\times\mathbb{R}_+^*;\frac{k}{h^2}\leq\frac{1}{2}\big\}$. The explicit three-point Euler scheme is stable under condition $\mathscr{S}$ for the norms $\|\cdot\|_{\infty,h}$, hence convergent for these norms.*

Proof. Let us choose the general form

$$\frac{1}{k}U^{j+1}-\frac{1}{k}U^j+A_hU^j=F^j,$$

for which we have consistency in the $\|\cdot\|_{\infty,h}$ norms and $G^j=F^j$. It can be rewritten as

$$U^{j+1}=\mathscr{A}_{k,h}U^j+kF^j,$$

where $\mathscr{A}_{k,h}=I-kA_h$. Therefore

$$\|U^{j+1}\|_{\infty,h}\leq|||\mathscr{A}_{k,h}|||_{\infty,h}\|U^j\|_{\infty,h}+k\|F^j\|_{\infty,h}.$$

Let us set $r=\frac{k}{h^2}$. By direct inspection, we see that

$$\mathscr{A}_{k,h}=\begin{pmatrix}1-2r & r & 0 & \cdots & 0\\ r & 1-2r & r & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & r\\ 0 & \cdots & 0 & r & 1-2r\end{pmatrix}.$$

It follows that

$$|||\mathscr{A}_{k,h}|||_{\infty,h} = |1-2r| + 2r = \begin{cases} 1 & \text{if } r \leq \frac{1}{2}, \\ 4r-1 & \text{if } r > \frac{1}{2}. \end{cases}$$

Therefore, if $r \leq \frac{1}{2}$, we have that

$$\begin{aligned} \|U^{j+1}\|_{\infty,h} &\leq \|U^j\|_{\infty,h} + k\|F^j\|_{\infty,h} \\ &\leq \|U^j\|_{\infty,h} + k \max_{n \leq T/k} \|F^n\|_{\infty,h}, \end{aligned}$$

therefore, iterating backwards, we obtain that for all j such that $j \leq \frac{T}{k}$,

$$\begin{aligned} \|U^j\|_{\infty,h} &\leq \|U^0\|_{\infty,h} + jk \max_{n \leq T/k} \|F^n\|_{\infty,h} \\ &\leq \|U^0\|_{\infty,h} + T \max_{n \leq T/k} \|F^n\|_{\infty,h}, \end{aligned}$$

hence the stability of the scheme for the norm ∞,h under condition $\mathscr{S}$. □

Remark 6.9.6 i) The above estimates are not sufficient to conclude that the scheme is not stable when $r > \frac{1}{2}$. However, numerical experiments with $r > \frac{1}{2}$ quickly show that the explicit Euler scheme is wildly non convergent for the ∞,h norm. Since it is consistent, this means it must be unstable. In particular, round-off errors are amplified exponentially fast.

ii) When $(h,k) \in \mathscr{S}$, we thus have $k \leq \frac{h^2}{2} \ll h$. For instance, if we want a modest amount of 1000 points in the space grid, then the time step must be smaller than $5\,10^{-7}$, *i.e.*, to compute up to a final time of $T = 1s$, we need at least $2\,10^6$ iterations in time. Such stability requirements can rapidly make the scheme too computationally expensive, in spite of its otherwise simplicity.

iii) The above example shows that it is fairly easy to give sufficient conditions of stability in the ∞,h norm for explicit schemes. □

Corollary 6.9.3 *When convergent, the explicit three-point Euler scheme has the error estimate*

$$\max_{i,j} |u_i^j - u(x_i,t_j)| \leq Ch^2.$$

Proof. This is a consequence of Corollary 6.9.2, Proposition 6.7.1, and the above remark on the set $\mathscr{S}$ which forces the time step to be much smaller than the space step. □

6.10 Stability for one time step schemes

We have seen that proving consistency is always a matter of combining several Taylor-Lagrange expansions together, which can be tedious but does not pose much difficulty in principle. Stability is another matter.

Let us consider a general scheme (6.20) with one time step, $l = 1$, $m = 0$ or $l = 0$, $m = 1$. We assume that the scheme can be rewritten as

$$U^{j+1} = \mathscr{A}U^j + kG^j \text{ or } U^{j+1} = \mathscr{A}U^j + kG^{j+1},$$

depending on whether the scheme is explicit or implicit. Actually, since G^j is an arbitrary source term, we are at liberty to rename G^{j+1} just G^j, it clearly changes nothing in terms of stability. Such a rewriting is possible in the case of both Euler methods.

The matrix $\mathscr{A}$ is called the *amplification matrix* of the scheme. It depends on k and h and it must not be forgotten that it is also of size $N \times N$, with $h = 1/(N+1)$. We give a first stability criterion.

Proposition 6.10.1 *A one time step general scheme* (6.20) *is stable for the family of norms* $\|\cdot\|_N$ *if and only if there exists a constant* $C(T)$ *depending only on* T *such that*[6]

$$\max_{j \leq T/k} |||\mathscr{A}^j|||_N \leq C(T). \tag{6.26}$$

Proof. Let us first assume that the scheme is stable. This means that there exist two constants $C_1(T)$ and $C_2(T)$ such that for any U_0 and G^j

$$\max_{j \leq T/k} \|U^j\|_N \leq C_1(T)\|U_0\|_N + C_2(T) \max_{j \leq T/k} \|G^j\|_N.$$

We take $G^j = 0$. In this case, $U^j = \mathscr{A}^j U_0$ and we have

$$\max_{j \leq T/k} \|\mathscr{A}^j U_0\|_N \leq C_1(T)\|U_0\|_N.$$

Since this is true for all $U_0 \in \mathbb{R}^N$, it follows that

$$\max_{j \leq T/k} |||\mathscr{A}^j|||_N \leq C_1(T).$$

[6] Beware of the notation: up to now U^j meant the jth vector in the sequence, but here $\mathscr{A}^j$ means the jth power of the matrix $\mathscr{A}$.

Conversely, assume that estimate (6.26) holds true. We can write

$$\begin{aligned} U^j &= \mathscr{A}U^{j-1} + kG^{j-1} \\ \mathscr{A}U^{j-1} &= \mathscr{A}^2U^{j-2} + k\mathscr{A}G^{j-2} \\ &\vdots \quad \vdots \\ \mathscr{A}^{j-1}U^1 &= \mathscr{A}^jU^0 + k\mathscr{A}^{j-1}G^0, \end{aligned}$$

so that summing these equations, we obtain

$$U^j = \mathscr{A}^jU^0 + k\sum_{n=0}^{j-1} \mathscr{A}^{j-n-1}G^n.$$

Therefore

$$\begin{aligned} \|U^j\|_N &\leq \|\mathscr{A}^jU^0\|_N + k\sum_{n=0}^{j-1} \|\mathscr{A}^{j-n-1}G^n\|_N \\ &\leq C(T)\|U^0\|_N + kC(T)\sum_{n=0}^{j-1} \|G^n\|_N \\ &\leq C(T)\|U^0\|_N + jkC(T)\max_{n\leq j}\|G^n\|_N \\ &\leq C(T)\|U^0\|_N + TC(T)\max_{n\leq T/k}\|G^n\|_N \end{aligned}$$

whenever $j \leq T/k$, hence the stability with $C_1(T) = C(T)$ and $C_2(T) = TC(T)$.□

The criterion given in Proposition 6.10.1 is not too practical in general, since the quantity $\max_j |||\mathscr{A}^j|||_N$ is not necessarily easy to estimate. Nonetheless, we have a sufficient condition as an immediate corollary.

Corollary 6.10.1 *If $|||\mathscr{A}|||_N \leq 1$, then the scheme is stable.*

Proof. An operator norm is submultiplicative, *i.e.*, $|||\mathscr{A}\mathscr{B}|||_N \leq |||\mathscr{A}|||_N|||\mathscr{B}|||_N$. Therefore, $|||\mathscr{A}^j|||_N \leq |||\mathscr{A}|||_N^j \leq 1$. □

This is what we did for the forward Euler scheme and the ∞, h norms. In the case of the $2, h$ norms, it is possible to be a little more precise.

Proposition 6.10.2 *If the amplification matrix $\mathscr{A}$ is normal, then the scheme is stable for the norms $\|\cdot\|_{2,h}$ if and only of there exists a constant $C'(T) \geq 0$ depending only on T such that*

$$\rho(\mathscr{A}) \leq 1 + C'(T)k. \tag{6.27}$$

Proof. We recall that a matrix $\mathscr{A}$ is said to be *normal* if $\mathscr{A}^T\mathscr{A} = \mathscr{A}\mathscr{A}^T$. In this case, $\rho(\mathscr{A}) = \rho(\mathscr{A}^T\mathscr{A})^{1/2} = |||\mathscr{A}|||_{2,h}$.

Let us first assume that there exists $C'(T) \geq 0$ such that $\rho(\mathscr{A}) \leq 1 + C'(T)k$. By hypothesis, $\mathscr{A}$ is normal, therefore $\mathscr{A}^j$ is also normal and $|||\mathscr{A}^j|||_{2,h} = \rho(\mathscr{A}^j) = \rho(\mathscr{A})^j$. Consequently, for all $j \leq T/k$,

$$|||\mathscr{A}^j|||_{2,h} \leq (1 + C'(T)k)^j \leq e^{C'(T)kj} \leq e^{C'(T)T},$$

and the constant $e^{C'(T)T}$ depends only on T. Therefore, the scheme is stable according to Proposition 6.10.1.

Conversely, assume that the scheme is stable. By Proposition 6.10.1 again, this implies that $\rho(\mathscr{A})^j \leq C(T)$ or $\rho(\mathscr{A}) \leq C(T)^{1/j}$ for all $j \leq T/k$. There are two cases. Either $C(T) \leq 1$ and thus $\rho(\mathscr{A}) \leq 1$ and we are done with $C'(T) = 0$, or $C(T) > 1$. In this case, we take $j = T/k$ so that

$$\rho(\mathscr{A}) \leq C(T)^{\frac{k}{T}} = e^{\frac{k}{T}\ln(C(T))},$$

with $\ln(C(T)) > 0$. This implies that the function $s \mapsto e^{s\ln(C(T))}$ is convex on $[0,1]$, which in turn implies that for all $s \in [0,1]$, $e^{s\ln(C(T))} \leq (1-s) + se^{\ln(C(T))} = 1 + s(C(T)-1)$. In particular, for $s = k/T$, we obtain

$$\rho(\mathscr{A}) \leq 1 + \frac{C(T)-1}{T}k,$$

hence estimate (6.27) with $C'(T) = \frac{C(T)-1}{T}$. □

Remark 6.10.1 It is important to stress again that the matrix $\mathscr{A}$ is a function of k and h, and so is its spectral radius. Therefore, the above estimates are by no means obvious.

Note that a sufficient condition for stability in the $2,h$ norm in the case of a normal amplification matrix, often used in practice, is thus that $\rho(\mathscr{A}) \leq 1$. This is particularly indicated if we are interested in the computation of long term behavior of the solution, *i.e.*, T large. Indeed, the less demanding condition (6.27) allows for exponential growth with T. □

Let us apply the above considerations to both Euler schemes and to the leapfrog scheme. We first need to determine the eigenvalues of the kind of tridiagonal matrices involved in these schemes.

Lemma 6.10.1 *Consider the $N \times N$ matrix*

$$A = \begin{pmatrix} a & b & 0 & \cdots & 0 \\ b & a & b & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & b & a & b \\ 0 & \cdots & 0 & b & a \end{pmatrix},$$

with $a, b \in \mathbb{R}$. The eigenvalues of A are given by

$$\lambda_n = a + 2b\cos\Big(\frac{n\pi}{N+1}\Big), n = 1, \ldots, N.$$

Proof. We have $A = aI + bB$ with $B = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 1 & 0 \end{pmatrix}$. Of course, $AV = \lambda V$ is equivalent to $BV = \frac{\lambda - a}{b} V$. It is thus sufficient to find the eigenvalues of B. For $n \in \mathbb{N}^*$, let $V_n \in \mathbb{R}^N \setminus \{0\}$ be the vector with coordinates $(V_n)_j = \sin\big(\frac{jn\pi}{N+1}\big)$, $j = 1, \ldots, N$. Then

$$BV_n = \begin{pmatrix} \sin\Big(\frac{2n\pi}{N+1}\Big) \\ \vdots \\ \sin\Big(\frac{(j-1)n\pi}{N+1}\Big) + \sin\Big(\frac{(j+1)n\pi}{N+1}\Big) \\ \vdots \\ \sin\Big(\frac{(N-1)n\pi}{N+1}\Big) \end{pmatrix}.$$

Each line is of the form $\sin\big(\frac{(j-1)n\pi}{N+1}\big) + \sin\big(\frac{(j+1)n\pi}{N+1}\big)$, even for $j = 1$ and $j = N$. Now,

$$\begin{aligned}(BV_n)_j &= \sin\Big(\frac{(j-1)n\pi}{N+1}\Big) + \sin\Big(\frac{(j+1)n\pi}{N+1}\Big) \\ &= 2\cos\Big(\frac{n\pi}{N+1}\Big)\sin\Big(\frac{jn\pi}{N+1}\Big) = 2\cos\Big(\frac{n\pi}{N+1}\Big)(V_n)_j,\end{aligned}$$

so that the numbers $2\cos\big(\frac{n\pi}{N+1}\big)$ are eigenvalues of B. For $n = 1, \ldots, N$, we thus obtain N distinct eigenvalues, hence we have found them all. □

Corollary 6.10.2 *The eigenvalues of A_h are*

$$\lambda_n = \frac{4}{h^2}\sin^2\Big(\frac{n\pi}{2(N+1)}\Big), n = 1, \ldots, N.$$

Proof. Apply the previous Lemma with $a = \frac{2}{h^2}$ and $b = -\frac{1}{h^2}$. □

We now return to the explicit Euler scheme.

Proposition 6.10.3 *Let $\mathscr{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$. The explicit three-point Euler scheme is stable for the norms $\|\cdot\|_{2,h}$ under condition $\mathscr{S}$ if and only if*

$$\mathscr{S} \subset \Big\{(h,k); \frac{k}{h^2}\cos^2\Big(\frac{\pi h}{2}\Big) \leq \frac{1}{2} + Ck\Big\}, \tag{6.28}$$

for some $C \geq 0$. In this case, it is convergent for these norms and we have the error estimate

$$\max_{j \leq T/k} \|U^j - S_h(u_{t_j})\|_{2,h} \leq Ch^2,$$

where C depends only on u and T.

Proof. We have $\mathscr{A} = I - kA_h$. It is a symmetric matrix, hence a normal matrix. We may thus apply Proposition 6.10.2. We have

$$\begin{aligned}\rho(\mathscr{A}) &= \max_{1 \leq n \leq N} \Big|1 - \frac{4k}{h^2}\sin^2\Big(\frac{n\pi}{2(N+1)}\Big)\Big| \\ &= \max\Big(1 - \frac{4k}{h^2}\sin^2\Big(\frac{\pi}{2(N+1)}\Big), \frac{4k}{h^2}\sin^2\Big(\frac{N\pi}{2(N+1)}\Big) - 1\Big).\end{aligned}$$

The first expression in the maximum is always between 0 and 1. We thus just need to consider the second expression. Condition (6.28) is then a simple rewriting of condition (6.27). □

Remark 6.10.2 The region $\frac{k}{h^2} \leq \frac{1}{2}$ is thus a stability region, as was the case for the ∞, h norm. Besides, convergence in the ∞, h norm implies convergence in the $2, h$ norm, so that nothing new is gained in this case. We know a little more about instability in the $2, h$ norm, which on the other hand does not directly imply instability in the ∞, h norm. However, using the converse part of the Lax theorem, if a consistent scheme is unstable in the $2, h$ norm, it is not convergent in the $2, h$ norm. Therefore, it is not convergent in the ∞, h norm, thus unstable for that same norm.

It should be noted that all these are theoretical considerations of convergence when $(h,k) \to (0,0)$. In practice, this is largely irrelevant since most of the time, only one value of h and k is used for actual computations. In this respect, it is nonetheless better to choose h and k such that $\frac{k}{h^2} \leq \frac{1}{2}$, since in this case, we are assured that $|||\mathscr{A}|||_{2,h} = \rho(\mathscr{A}) < 1$, which is numerically a good thing. □

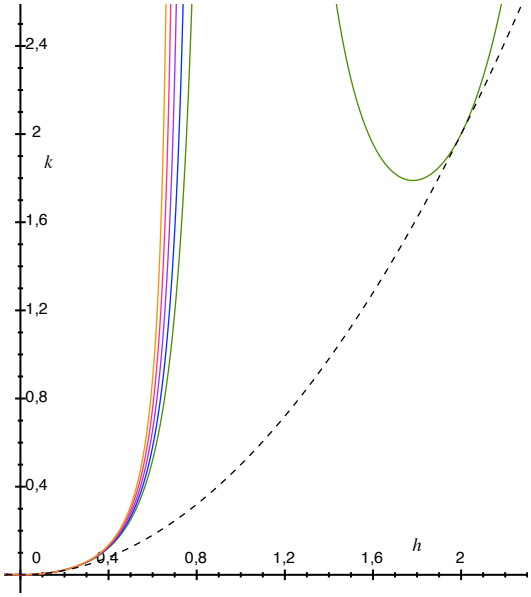


Figure 9. The boundaries of a few stability regions ($C = 0, 0.1, 0.2, 0.3, 0.4$) and of the "safe" region $k \leq h^2/2$ dashed. They are all tangent at $(0,0)$.

We see from the above picture that there is no practical difference between the different stability regions where it counts, that is to say in a neighborhood of $(0,0)$. Let us now turn to the implicit Euler scheme.

Proposition 6.10.4 *The implicit three-point Euler scheme is unconditionally stable for the norms $\|\cdot\|_{2,h}$. It is convergent for these norms and we have the error estimate*

$$\max_{j \leq T/k} \|U^j - S_h(u_{t_j})\|_{2,h} \leq C(h^2 + k),$$

where C depends only on u and T.

Proof. We have $\mathscr{A} = (I + kA_h)^{-1}$, which is symmetric, hence normal. Its eigenvalues are

$$\lambda_n = \frac{1}{1 + \frac{4k}{h^2} \sin^2\left(\frac{n\pi}{2(N+1)}\right)}, n = 1, \ldots, N,$$

and are all between 0 and 1. Hence $\rho(\mathscr{A}) < 1$ and the scheme is unconditionally stable. □

Remark 6.10.3 We see here the great advantage of the implicit Euler scheme over the explicit Euler scheme. The number of time iterations to reach a given time T is not constrained by the space step. □

We finally deal with the leapfrog scheme. The leapfrog scheme is a two time step scheme. In order to apply the above stability result, we need to rewrite it as a single time step scheme. It suffices to double the dimension. Indeed, if we set

$$V^j = \begin{pmatrix} U^j \\ U^{j-1} \end{pmatrix} \in \mathbb{R}^{2N},$$

then the leapfrog scheme

$$U^{j+1} = U^{j-1} - 2kA_hU^j + 2kF^j$$

becomes

$$\begin{aligned} V^{j+1} = \begin{pmatrix} U^{j+1} \\ U^j \end{pmatrix} &= \begin{pmatrix} U^{j-1} - 2kA_hU^j \\ U^j \end{pmatrix} + \begin{pmatrix} 2kF^j \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -2kA_h & I \\ I & 0 \end{pmatrix} \begin{pmatrix} U^j \\ U^{j-1} \end{pmatrix} + \begin{pmatrix} 2kF^j \\ 0 \end{pmatrix} \\ &= \mathscr{B}V^j + kG^j \end{aligned}$$

with a $2N \times 2N$ symmetric amplification matrix $\mathscr{B}$ and $G^j \in \mathbb{R}^{2N}$. We need to find the spectral radius of this matrix.

Lemma 6.10.2 *Let C be a $N \times N$ complex matrix and B the $2N \times 2N$ complex matrix defined by blocks as*

$$B = \begin{pmatrix} C & I \\ I & 0 \end{pmatrix}.$$

If $\lambda \in \mathbb{C}$ is an eigenvalue of B, then $\lambda \neq 0$ and $\lambda - \frac{1}{\lambda}$ is an eigenvalue of C. Conversely, if $\mu \in \mathbb{C}$ is an eigenvalue of C, then there exists an eigenvalue λ of B such that $\mu = \lambda - \frac{1}{\lambda}$.

Proof. Let $\lambda \in \mathbb{C}$ be such that there exists a vector Y in $\mathbb{C}^{2N}$, $0 \neq Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, such that $BY = \lambda Y$. Using the block structure of B, we see that this is equivalent to

$$\begin{cases} CY_1 + Y_2 = \lambda Y_1, \\ Y_1 = \lambda Y_2. \end{cases}$$

If $Y_1 = 0$, the first equation implies that $Y_2 = 0$, which is impossible. Thus $Y_1 \neq 0$, which implies $\lambda \neq 0$ by the second equation. We may thus divide by λ so that $Y_2 = \frac{1}{\lambda}Y_1$ and replacing in the first equation $CY_1 = \left(\lambda - \frac{1}{\lambda}\right)Y_1$. Since we have already seen that $Y_1 \neq 0$, this implies that $\lambda - \frac{1}{\lambda}$ is an eigenvalue of C.

Conversely, let $\mu \in \mathbb{C}$ be an eigenvalue of C with eigenvector $0 \neq Y_1 \in \mathbb{C}^N$. The polynomial $X^2 - \mu X - 1$ has two roots in $\mathbb{C}$, which are nonzero since their product is -1. Let λ be one of these roots. Dividing by λ, we see that $\lambda - \mu - \frac{1}{\lambda} = 0$, hence $\mu = \lambda - \frac{1}{\lambda}$. Furthermore

$$B \begin{pmatrix} Y_1 \\ \frac{1}{\lambda} Y_1 \end{pmatrix} = \begin{pmatrix} CY_1 + \frac{1}{\lambda} Y_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} (\mu + \frac{1}{\lambda}) Y_1 \\ Y_1 \end{pmatrix} = \lambda \begin{pmatrix} Y_1 \\ \frac{1}{\lambda} Y_1 \end{pmatrix},$$

so that λ is an eigenvalue of B. □

Remark 6.10.4 If λ is an eigenvalue of B, then $-\frac{1}{\lambda}$ is also an eigenvalue of B. This pair corresponds to the same eigenvalue μ of C. □

Let us now apply this to the leapfrog scheme.

Proposition 6.10.5 *The leapfrog scheme is unstable for the norms $\|\cdot\|_{2,h}$, hence not convergent for these norms.*

Proof. The matrix $\mathscr{B}$ is symmetric, hence normal. We may thus apply Proposition 6.10.2.

The eigenvalues of the matrix $-2kA_h$ are

$$\mu_n = -\frac{8k}{h^2} \sin^2\Big(\frac{n\pi}{2(N+1)}\Big), n = 1, \ldots, N,$$

and those of the matrix $\mathscr{B}$

$$\lambda_n^{\pm} = \frac{\mu_n \pm \sqrt{\mu_n^2 + 4}}{2}$$

according to Lemma 6.10.2. In particular, for $n = N$, we have

$$\sin^2\Big(\frac{N\pi}{2(N+1)}\Big) = \cos^2\Big(\frac{\pi}{2(N+1)}\Big) \geq \frac{1}{2}$$

since $\frac{\pi}{2(N+1)} \leq \frac{\pi}{4}$. Therefore

$$-\mu_N \geq \frac{4k}{h^2}.$$

It follows that

$$\rho(\mathscr{B}) \geq \left| \frac{\mu_N - \sqrt{\mu_N^2 + 4}}{2} \right| = \frac{\sqrt{\mu_N^2 + 4} - \mu_N}{2} \geq \frac{2 + \frac{4k}{h^2}}{2} = 1 + \frac{2k}{h^2}.$$

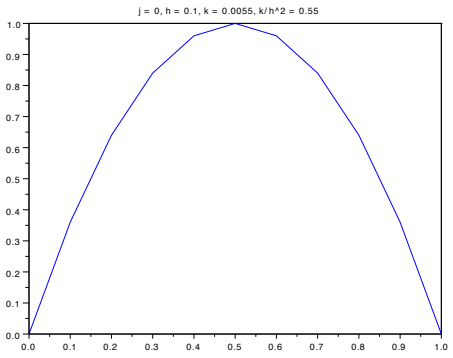
Consequently, there is no constant C such that $\rho(\mathscr{B}) \leq 1 + Ck$. Indeed, assume there is such a constant, for (h,k) in some region $\mathscr{S}$. Then we would have $\frac{2}{h^2} \leq C$, which precludes $h \to 0$. This is inconsistent with $(0,0) \in \bar{\mathscr{S}}$. □

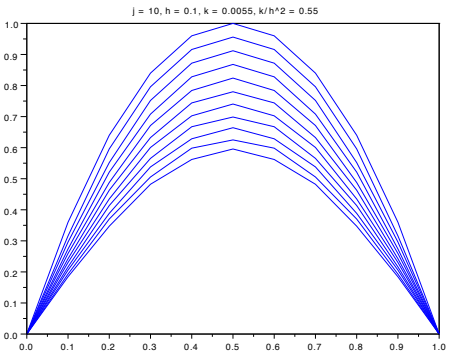
Remark 6.10.5 The leapfrog scheme is thus not usable in practice. Numerical experiments show that it diverges very rapidly. We remark in addition that

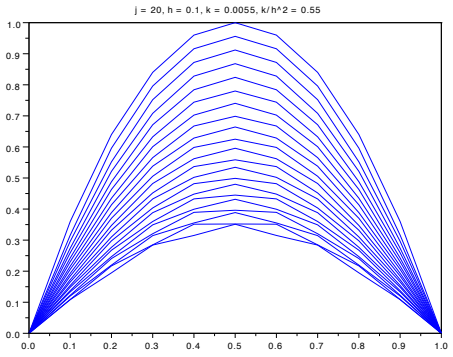
$$\rho(\mathscr{B}^{M+1}) \geq \left(1+\frac{2k}{h^2}\right)^{M+1} \geq 1+\frac{2(M+1)k}{h^2} = 1+\frac{2T}{h^2} \to +\infty \text{ when } h \to 0.$$

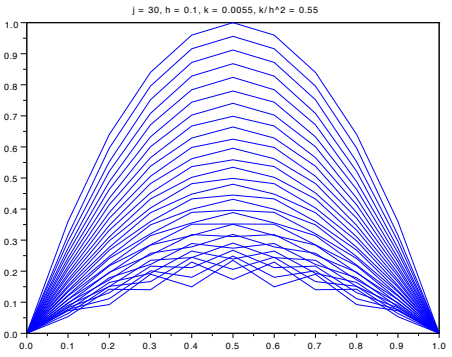
The same is true for any number of iterations needed to reach a fixed time $t > 0$. □

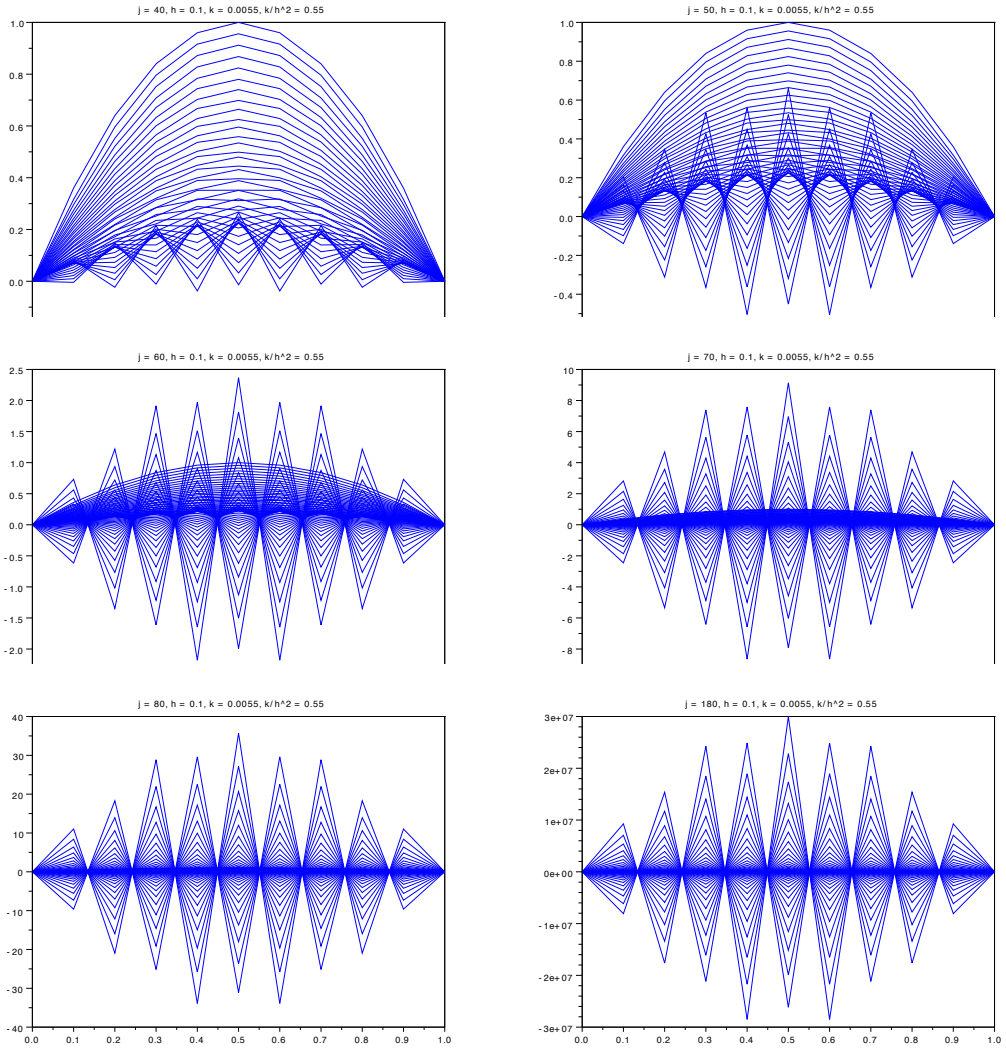
Below is a sequence of plots corresponding to the forward Euler method applied to $u_0(x) = 4x(1-x)$ in a case when $\frac{k}{h^2} = 0.55 > \frac{1}{2}$. For each value of j, the piecewise affine interpolate of the values u_i^j is plotted. The onset of instability occurs between $j = 10$ and $j = 20$ and then only gets worse. Notice the scale on the last plot ($j = 180$).











6.11 Stability via the Fourier transform

Stability for the $2, h$ norms is closely related to the spectral radius of the amplification matrix, at least when the latter is normal. Unfortunately, it is not always easy to compute the eigenvalues of a matrix. We now present an alternate way using the Fourier transform, which is not directly applicable to the previously introduced schemes—in fact it applies to a different family of objects—but that still gives stability information in a much more workable fashion.

We thus now work with the heat equation on $\mathbb{R}$, without boundary conditions.

For definiteness, let us consider the forward Euler scheme

$$\begin{cases} \dfrac{u_i^{j+1}-u_i^j}{k} - \dfrac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} = 0, i \in \mathbb{Z}, \\ u_i^0 = u_0(ih), i \in \mathbb{Z}. \end{cases} \tag{6.29}$$

If we assume that $(u_i^0)_{i\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, *i.e.* that $\sum_{i\in\mathbb{Z}} |u_i^0|^2 < +\infty$, then it is quite clear that $(u_i^j)_{i\in\mathbb{Z}}$ is well defined and belongs to $\ell^2(\mathbb{Z})$ for all j. We equip $\ell^2(\mathbb{Z})$ with the norm

$$\|(v_i)_{i\in\mathbb{Z}}\|_{2,h} = \sqrt{h}\Big(\sum_{i\in\mathbb{Z}} |v_i|^2\Big)^{1/2}$$

for which it is a Hilbert space, using the same notation as in the bounded interval case. Of course, the forward Euler scheme is also defined on other spaces of $\mathbb{Z}$-indexed sequences, but we concentrate here on ℓ^2. It should be noted that such schemes are not implementable in practice, since they involve an infinite number of unknowns. Their interest is purely theoretical.

Instead of working directly with the above discrete scheme, we introduce a *semi-discrete* version of it. In a semi-discrete scheme, only time is fully discretized. Space is only semi-discretized in the sense that it remains continuous even though we retain the space step h. We thus consider sequences of functions $u^j : \mathbb{R} \to \mathbb{R}$ which are such that u^j is supposed to be an approximation of $x \mapsto u(x,t_j)$.

The semi-discrete version of the forward Euler scheme is as follows:

$$\begin{cases} \dfrac{u^{j+1}(x)-u^j(x)}{k} - \dfrac{u^j(x+h) - 2u^j(x) + u^j(x-h)}{h^2} = 0, \\ u^0(x) = u_{0,h}(x). \end{cases} \tag{6.30}$$

where $u_{0,h}$ is some approximation of u_0. So the idea is to use the differential quotient on which the discrete scheme is based to approximate the space derivative, and the usual discrete difference quotient for the time derivative. This way, any discrete scheme admits a semi-discrete version.

A good functional setting for this is for example $L^2(\mathbb{R})$. Indeed, if $u_{0,h} \in L^p(\mathbb{R})$, then clearly, u^j is well defined and belongs to $L^p(\mathbb{R})$. In effect, if $u_{0,h} \in L^2(\mathbb{R})$, then we can write $u^{j+1} = G(u^j)$ where G is the continuous linear operator in $\mathscr{L}(L^2(\mathbb{R}), L^2(\mathbb{R}))$ defined by

$$Gv(x) = v(x) + \frac{k}{h^2}(v(x+h) - 2v(x) + v(x-h)), \tag{6.31}$$

(notice that the right-hand side of the equation is 0), or equivalently $G = (1 - \frac{2k}{h^2})I + \frac{k}{h^2}(\tau_h + \tau_{-h})$, where τ_s denotes the operator of translation by s, $\tau_s u(x) = u(x+s)$.

Therefore, $u^j = G^j(u_{0,h})$ and the properties of the scheme are the properties of the iterates of the operator G, provided $u_{0,h}$ remains bounded.

Before continuing further on the general study of semi-discrete schemes, let us discuss the relationship between the fully discrete and semi-discrete points of view. We need to associate a function in $L^2(\mathbb{R})$ with each sequence of numbers in $\ell^2(\mathbb{Z})$.

Proposition 6.11.1 *For all $v \in \ell^2(\mathbb{Z})$, we define a piecewise constant interpolation $I_h v$ by*

$$\forall i \in \mathbb{Z}, \forall x \in \Big]x_i - \frac{h}{2}, x_i + \frac{h}{2}\Big[, \quad I_h v(x) = v_i. \tag{6.32}$$

The interpolation operator I_h is an isometry between $\ell^2(\mathbb{Z})$ equipped with the $\|\cdot\|_{2,h}$ norm and $L^2(\mathbb{R})$.

Proof. Indeed,

$$\|I_h v\|^2_{L^2(\mathbb{R})} = \int_{\mathbb{R}} I_h v(x)^2\,dx = \sum_{i\in\mathbb{Z}} \int_{x_i-\frac{h}{2}}^{x_i+\frac{h}{2}} I_h v(x)^2\,dx = \sum_{i\in\mathbb{Z}} h v_i^2 = \|v\|^2_{2,h}, \tag{6.33}$$

and the proof is complete. □

Let us now see that the discrete scheme and the semi-discrete scheme are equivalent when the initial data of the semi-discrete scheme is in the range of I_h.

Proposition 6.11.2 *Let us be given $(u_i^0)_{i\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. If $u_{0,h} = I_h\big((u_i^0)_{i\in\mathbb{Z}}\big)$, then $u^j = I_h\big((u_i^j)_{i\in\mathbb{Z}}\big)$ for all $j \in \mathbb{N}$.*

Proof. We prove this by induction on j. The statement is true for $j = 0$ by hypothesis. Let us thus assume that $u^j = I_h\big((u_i^j)_{i\in\mathbb{Z}}\big)$. This means that for all $x \in \big]x_i - \frac{h}{2}, x_i + \frac{h}{2}\big[$, we have $u^j(x) = u_i^j$. Therefore, in view of (6.31), for the same values of x, we have

$$\begin{aligned} Gu^j(x) &= u^j(x) + \frac{k}{h^2}(u^j(x+h) - 2u^j(x) + u^j(x-h)) \\ &= u_i^j + \frac{k}{h^2}(u_{i+1}^j - 2u_i^j + u_{i-1}^j) = u_i^{j+1}, \end{aligned}$$

so that $u^{j+1} = Gu^j = I_h\big((u_i^{j+1})_{i\in\mathbb{Z}}\big)$. □

So the idea is that, if we start the semi-discrete scheme with an initial data constructed by piecewise interpolation from the discrete scheme, the semi-discrete scheme will construct exactly the same values as the discrete scheme. The advantage is that the semi-discrete scheme works for much more general initial data, which in turn makes the study of stability considerably easier.

Definition 6.11.1 *We say that the semi-discrete scheme is stable in L^2 if there exists a constant $C(T)$ such that*

$$\max_{j\le T/k} \|u^j\|_{L^2(\mathbb{R})} \le C(T)\|u_{0,h}\|_{L^2(\mathbb{R})}, \tag{6.34}$$

for all $u_{0,h} \in L^2(\mathbb{R})$.

Here $u_{0,h}$ is no longer to be thought of as some approximation of u_0. Clearly, this is equivalent to $\|G^j\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))}$ being bounded independently of j, h and k.[7] In view of Propositions 6.11.1 and 6.11.2, stability of the semi-discrete scheme implies stability of the discrete scheme in the $\|\cdot\|_{2,h}$ norms, hence the interest of the approach.

The reason for singling out L^2 among all L^p spaces is that the Fourier transform is an isometry on L^2. Let us briefly recall a few facts about the Fourier transform. When $u \in L^1(\mathbb{R})$, the Fourier transform of u is defined by

$$\widehat{u}(\xi) = \mathscr{F}u(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-ix\xi} u(x)\,dx.$$

The function $\widehat{u}$ is continuous and tends to 0 at infinity. When $u \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, it can be shown that it also belongs to $L^2(\mathbb{R})$ and that $\|\widehat{u}\|_{L^2(\mathbb{R})} = \|u\|_{L^2(\mathbb{R})}$, which is called the Plancherel formula. Thus the Fourier transform extends as an isometry to the whole of L^2 by density of $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ in $L^2(\mathbb{R})$ (but not by the simple Lebesgue integral formula above, which makes no sense in the L^2 context).

The feature of the Fourier transform that makes it so useful here, in addition to being an isometry, is that it transforms translations into multiplications by exponentials. More precisely, if $u \in L^1(\mathbb{R})$ and $s \in \mathbb{R}$, then

$$\widehat{\tau_s u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-ix\xi} u(x+s)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i(y-s)\xi} u(y)\,dy = e^{is\xi}\widehat{u}(\xi),$$

and the equality between the two ends remains true for any $u \in L^2(\mathbb{R})$ by density.

Proposition 6.11.3 *Let $a(\xi) = 1 - \frac{4k}{h^2}\sin^2\left(\frac{h\xi}{2}\right)$. Then we have*

$$\|G^j\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))} = \sup_{\xi\in\mathbb{R}} |a(\xi)|^j.$$

Proof. We apply the Fourier transform to the semi-discrete scheme (6.30). This yields

$$\frac{\widehat{u^{j+1}}(x) - \widehat{u^j}(x)}{k} - \frac{e^{ih\xi} - 2 + e^{-ih\xi}}{h^2}\widehat{u^j}(x) = 0,$$

[7] Here again, G depends on h and k even though the notation does not make it plain.

or

$$\widehat{u^{j+1}}(x) = \widehat{u^j}(x) + \frac{2k}{h^2}(\cos(h\xi) - 1)\widehat{u^j}(x),$$

or again

$$\widehat{u^{j+1}}(x) = a(\xi)\widehat{u^j}(x).$$

Iterating this relation, we obtain

$$\mathscr{F}(G^j u_0)(\xi) = \widehat{u^j}(x) = a(\xi)^j \widehat{u_0}(x).^8$$

Let now M be a Fourier multiplier operator, *i.e.*, an operator such that

$$\mathscr{F}(Mu) = m\widehat{u}$$

with $m \in L^\infty(\mathbb{R})$, which is the case of G^j above. Let us show that

$$\|M\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))} = \|m\|_{L^\infty(\mathbb{R})}.$$

First of all, for all $u \in L^2(\mathbb{R})$, we have

$$\begin{aligned}\|Mu\|^2_{L^2(\mathbb{R})} = \|\mathscr{F}(Mu)\|^2_{L^2(\mathbb{R})} &= \int_{\mathbb{R}} |m(\xi)|^2 |\widehat{u}(\xi)|^2 \, d\xi \\ &\leq \|m\|^2_{L^\infty(\mathbb{R})} \|\widehat{u}\|^2_{L^2(\mathbb{R})} = \|m\|^2_{L^\infty(\mathbb{R})} \|u\|^2_{L^2(\mathbb{R})},\end{aligned}$$

so that

$$\|M\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))} \leq \|m\|_{L^\infty(\mathbb{R})}.$$

Next, let $\varepsilon > 0$ and $A \subset \mathbb{R}$ a set of strictly positive, finite measure such that $|m(\xi)| \geq \|m\|_{L^\infty(\mathbb{R})} - \varepsilon \geq 0$ on A. We take $u \in L^2(\mathbb{R})$ such that $\widehat{u} = (\operatorname{meas} A)^{-1/2} \mathbf{1}_A$. Then $\|u\|_{L^2(\mathbb{R})} = 1$ and

$$\begin{aligned}\|Mu\|^2_{L^2(\mathbb{R})} = \int_{\mathbb{R}} |m(\xi)|^2 |\widehat{u}(\xi)|^2 \, d\xi &= \frac{1}{\operatorname{meas} A} \int_A |m(\xi)|^2 \, d\xi \\ &\geq \frac{(\|m\|_{L^\infty(\mathbb{R})} - \varepsilon)^2}{\operatorname{meas} A} \int_A d\xi = (\|m\|_{L^\infty(\mathbb{R})} - \varepsilon)^2.\end{aligned}$$

Therefore

$$\|M\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))} \geq \|m\|_{L^\infty(\mathbb{R})} - \varepsilon$$

for all $\varepsilon > 0$, and the proposition is proved. □

[8] Again, beware of the notation: u^j is the jth function in the sequence, whereas G^j is the jth iterate of the operator G and a^j is the function a to the power j.

Proposition 6.11.4 *The forward Euler semi-discrete scheme is stable in L^2 if $\frac{k}{h^2} \leq \frac{1}{2}$ and unstable if $\frac{k}{h^2} \geq \lambda > \frac{1}{2}$.*

Proof. We have $a(\xi) = 1 - \frac{4k}{h^2}\sin^2\left(\frac{h\xi}{2}\right) \leq 1$ for all $\xi \in \mathbb{R}$ and $a(0) = 1$. On the other hand, the minimum of $a(\xi)$ is attained for $\frac{h\xi}{2} = \frac{\pi}{2} + k\pi$ and its minimum value is $1 - \frac{4k}{h^2}$. Therefore

$$\sup_{\xi\in\mathbb{R}} |a(\xi)| = \max\left(1, \left|1 - \frac{4k}{h^2}\right|\right).$$

Consequently, if $\frac{k}{h^2} \leq \frac{1}{2}$, then $\sup_{\xi\in\mathbb{R}} |a(\xi)| = 1$ so that $\|G^j\|_{\mathscr{L}(L^2(\mathbb{R}),L^2(\mathbb{R}))} = 1$ and the scheme is stable in L^2.

If, on the other hand, $\frac{k}{h^2} \geq \lambda > \frac{1}{2}$, then

$$\sup_{\xi\in\mathbb{R}} |a(\xi)|^j \geq (4\lambda - 1)^j,$$

so that

$$\max_{j\leq T/k} \sup_{\xi\in\mathbb{R}} |a(\xi)|^j \geq (4\lambda - 1)^{T/k} \to +\infty \text{ when } k \to 0,$$

hence the scheme is unstable. □

Corollary 6.11.1 *The forward Euler discrete scheme is stable in the $\|\cdot\|_{2,h}$ norms if $\frac{k}{h^2} \leq \frac{1}{2}$.*

We can apply the same philosophy to a general single time step finite difference scheme and obtain corresponding semi-discrete schemes which are of the form $\widehat{u^{j+1}}(\xi) = a(\xi)\widehat{u^j}(\xi)$, $\widehat{u^0}$ given, in Fourier space. The function a, which depends on h and k as parameters, is called the *amplification coefficient* of the scheme.

Using the same kind of arguments as those used with matrices, it is possible to prove that a scheme is stable in L^2 if and only if there exists a positive constant C that depends only on T such that $|a(\xi)| \leq 1 + Ck$ for all ξ.

Definition 6.11.2 *We say that a semi-discrete scheme is* stable in the sense of von Neumann *if* $\sup_{\xi\in\mathbb{R}} |a(\xi)| \leq 1$.

Clearly, stability in the sense of von Neumann implies stability in L^2 for all T and uniformly with respect to T. It is thus a sufficient condition of stability for both semi-discrete and discrete schemes. Obviously, computations in Fourier space are much easier than evaluations of spectral radii.

We now consider the example of a family of schemes, collectively known as *the θ-scheme*. Let us be given a number $\theta \in [0,1]$. The discrete version of the θ-scheme is as follows.

$$\frac{u_i^{j+1}-u_i^j}{k} - \theta\frac{u_{i+1}^{j+1}-2u_i^{j+1}+u_{i-1}^{j+1}}{h^2} - (1-\theta)\frac{u_{i+1}^{j}-2u_i^{j}+u_{i-1}^{j}}{h^2} = \theta f_i^{j+1} + (1-\theta) f_i^j, \quad (6.35)$$

with boundary and initial conditions. The θ-scheme is thus a weighted average of the explicit Euler scheme ($\theta = 0$) and the implicit Euler scheme ($\theta = 1$). It is implicit as soon as $\theta > 0$.

The semi-discrete version of the θ-scheme (with 0 right-hand side) is then

$$\frac{u^{j+1}(x)-u^j(x)}{k} - \theta\frac{u^{j+1}(x+h)-2u^{j+1}(x)+u^{j+1}(x-h)}{h^2} - (1-\theta)\frac{u^{j}(x+h)-2u^{j}(x)+u^{j}(x-h)}{h^2} = 0. \quad (6.36)$$

Note that the scheme is implicit for $\theta > 0$ and it is not clear that such an implicit scheme is even well-defined on $L^2(\mathbb{R})$. The Fourier transform is again the key here. Indeed, in Fourier space, we have

$$\frac{\widehat{u^{j+1}}(\xi)-\widehat{u^j}(\xi)}{k} - \theta\frac{e^{ih\xi}-2+e^{-ih\xi}}{h^2}\widehat{u^{j+1}}(\xi) - (1-\theta)\frac{e^{ih\xi}-2+e^{-ih\xi}}{h^2}\widehat{u^j}(\xi) = 0,$$

which boils down to

$$\Big(1+\theta\frac{4k}{h^2}\sin^2\Big(\frac{h\xi}{2}\Big)\Big)\widehat{u^{j+1}}(\xi) = \Big(1-(1-\theta)\frac{4k}{h^2}\sin^2\Big(\frac{h\xi}{2}\Big)\Big)\widehat{u^j}(\xi).$$

Now we see that $1+\theta\frac{4k}{h^2}\sin^2\big(\frac{h\xi}{2}\big) \geq 1$, hence, its inverse is in L^∞ and the scheme can be rewritten as a Fourier multiplier operator with amplification coefficient

$$a(\xi) = \frac{1-(1-\theta)\frac{4k}{h^2}\sin^2\big(\frac{h\xi}{2}\big)}{1+\theta\frac{4k}{h^2}\sin^2\big(\frac{h\xi}{2}\big)} \in L^\infty(\mathbb{R}).$$

Clearly, $a(\xi) \leq 1$ for all $\xi \in \mathbb{R}$. Stability in the sense of von Neumann thus depends on whether or not we have $a(\xi) \geq -1$ for all ξ.

Proposition 6.11.5 *If $\theta \geq \frac{1}{2}$, then the θ-scheme is unconditionally stable in the sense von Neumann. If $\theta < \frac{1}{2}$, it is stable in the sense of von Neumann under the condition $\frac{k}{h^2} \leq \frac{1}{2(1-2\theta)}$.*

Proof. After a little bit of computation, it can be checked that $a(\xi) \geq -1$ if and only if $1 + (2\theta - 1)\frac{2k}{h^2}\sin^2\left(\frac{h\xi}{2}\right) \geq 0$, hence the result. □

Returning to the discrete version of the scheme, what about consistency and order for the θ-scheme?

Proposition 6.11.6 *The θ-scheme is of order 1 in time and 2 in space for $\theta \neq \frac{1}{2}$ in the ∞, h norm, and of order 2 in time and 2 in space for $\theta = \frac{1}{2}$.*

Proof. Let us list the results of the application of Taylor-Lagrange expansions to the various terms, without writing the remainders explicitly. For the time derivative, we have

$$\begin{aligned}\frac{u(x_i,t_{j+1}) - u(x_i,t_j)}{k} &= \frac{\partial u}{\partial t}(x_i,t_j) + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,t_j) + O(k^2) \\ &= \frac{\partial u}{\partial t}(x_i,t_{j+1}) - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,t_{j+1}) + O(k^2)\end{aligned}$$

For the space derivatives, we obtain

$$\frac{u(x_{i+1},t_j) - 2u(x_i,t_j) + u(x_{i-1},t_j)}{h^2} = \frac{\partial^2 u}{\partial x^2}(x_i,t_j) + O(h^2)$$

and

$$\frac{u(x_{i+1},t_{j+1}) - 2u(x_i,t_{j+1}) + u(x_{i-1},t_{j+1})}{h^2} = \frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1}) + O(h^2).$$

Therefore, combining these relations together, we have

$$\begin{aligned}\varepsilon(u)^j = \theta\Big(\frac{\partial u}{\partial t}(x_i,t_{j+1}) - \frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1})\Big) + (1-\theta)\Big(\frac{\partial u}{\partial t}(x_i,t_j) - \frac{\partial^2 u}{\partial x^2}(x_i,t_j)\Big) \\ - \theta\frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,t_{j+1}) + (1-\theta)\frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i,t_j) + O(k^2) + O(h^2) \\ - \theta f_i^{j+1} - (1-\theta) f_i^j.\end{aligned}$$

Now we can write

$$\frac{\partial^2 u}{\partial t^2}(x_i,t_{j+1}) = \frac{\partial^2 u}{\partial t^2}(x_i,t_j) + O(k).$$

Canceling all cancelable terms, we thus obtain

$$\varepsilon(u)^j = k\Big(\frac{1}{2} - \theta\Big)\frac{\partial^2 u}{\partial t^2}(x_i,t_j) + O(k^2) + O(h^2),$$

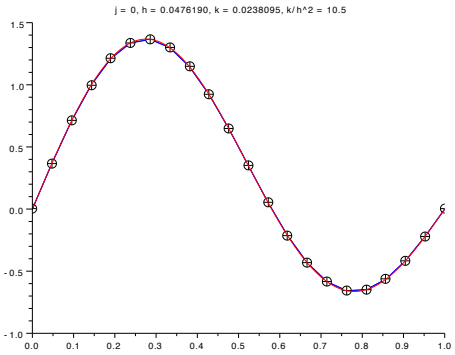
and the result follows. □

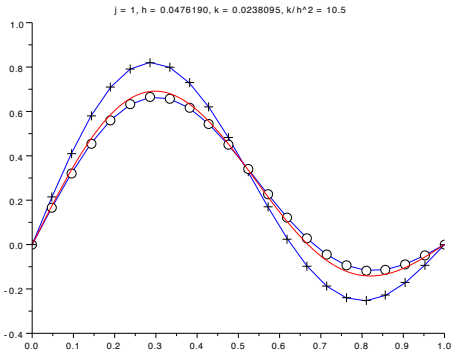
Remark 6.11.1 The θ-scheme for $\theta = \frac{1}{2}$ thus appears to be particularly attractive: it is unconditionally (von Neumann) stable and of order 2 in time and space. Of course, we have not really proved here that the discrete scheme on a bounded interval is actually unconditionally stable for the ∞, h or $2, h$ norms, but this is nonetheless true. It is called the *Crank-Nicolson scheme*. We rewrite it here in full

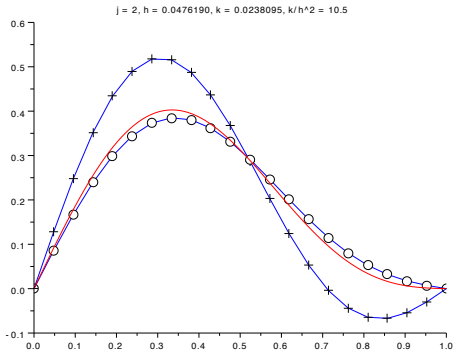
$$\begin{cases} \dfrac{u_i^{j+1} - u_i^j}{k} - \dfrac{1}{2h^2}\left(u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1} + u_{i+1}^j - 2u_i^j + u_{i-1}^j\right) = \dfrac{1}{2}\left(f_i^{j+1} + f_i^j\right) \\ u^0 = u_0. \end{cases} \tag{6.37}$$

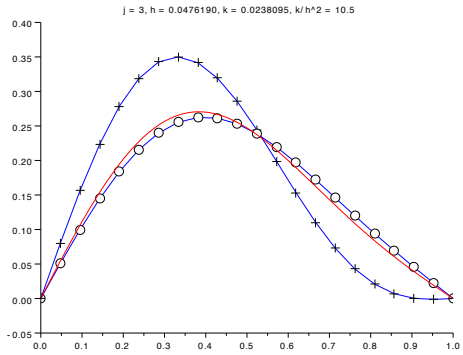
The scheme is implicit, with the same computational cost as the backward Euler scheme, since the evaluation of U^{j+1} entails solving a tridiagonal linear system with a very similar matrix. □

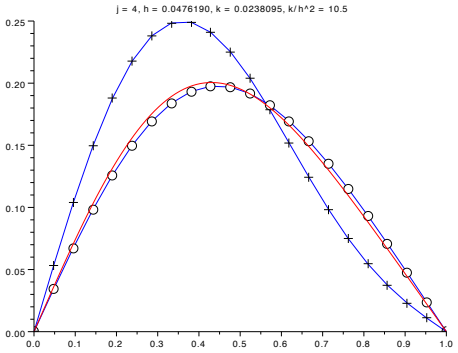
For purposes of comparison, we plot below the results of the backward Euler scheme, the Crank-Nicolson scheme and the exact solution, for various values of j on the same graphs, for the initial data $u_0(x) = \sin(\pi x)/2 + \sin(2\pi x)$ and the same discretization parameters h and k. The Euler scheme is indicated with $+$ marks (and linearly interpolated in blue), the Crank-Nicolson scheme with $\circ$ marks and the exact solution with a solid red line. The scale varies from plot to plot.

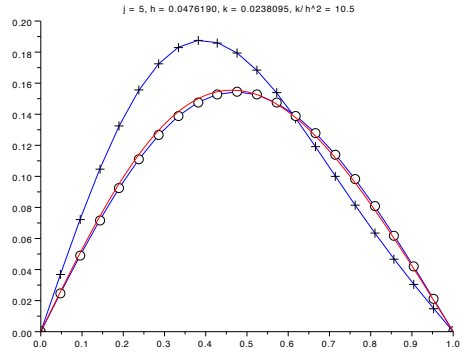


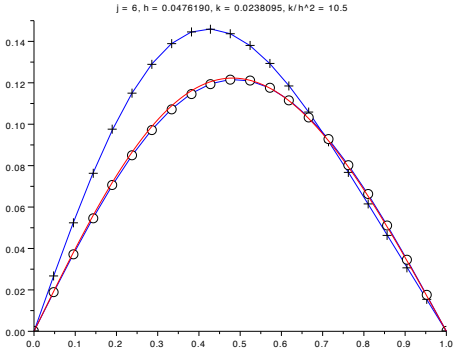


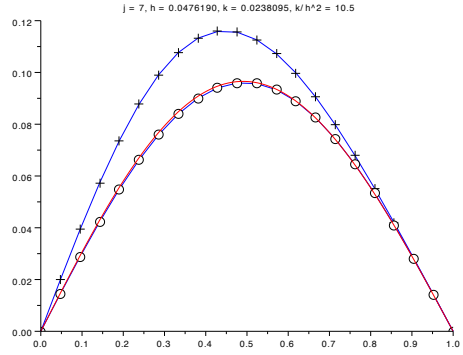

j = 2, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5

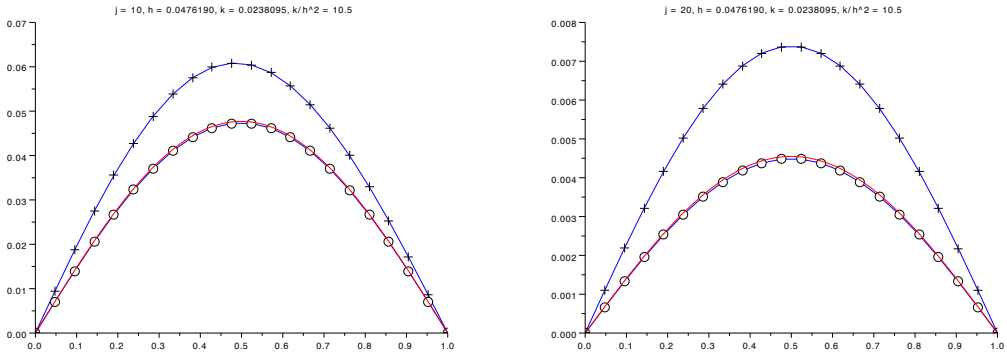

j = 3, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5


j = 4, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5


j = 5, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5


j = 6, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5


j = 7, h = 0.0476190, k = 0.0238095, k/h^2 = 10.5



Both schemes are stable and the higher order, hence better accuracy, of the Crank-Nicolson scheme is clearly visible for this particular initial data.