A Master's course: An introduction to the numerical analysis of partial differential equations

Author: Dr. Bradji, Abdallah

Provisional home Page : http://www.cmi.univ-mrs.fr/~bradji

- This document is not finished yet: last update Wednesday 18th August 2010
- This is a first draft. Would be kind from the reader if could provide me with any mistake may be found in this document, suggestion,... .

Remarks on the document

Below, I quote some remarks to be taken in consideration:

1. The stability of the finite difference scheme of [227]–[231], in Section 8, is proven using an idea for [GOD 77, Pages 268–269]. I'm feeling that such stability could be proved in a simpler way using Lemmata 8.1 and 8.2.
2. I enjoyed very well some comments quoted by [GOD 77, Pages 239–253] about the regularity required to get the convergence of the finite difference schemes. I quote here some of these useful remarks in Section 10. I'm so interested with the question of the regularity assumption on the exact solution which is required in the numerical methods. At least, for two reason make me so interested with the wonderful question of regularity:
   - recently i'm interested with the numerical approximation of hyperbolic equation in which the exact solution is not so smooth,
   - is possible to get higher order approximations with some basic regularity assumptions on the exact solution of the equation to be resolved.
3. there is a second item not written yet in Section 10. This item consists of the second issue to manage with the numerical approximation of non smooth data. Is not so clear yet for me ...
4. the first example, it is the Bürgers equation, not fnished yet.

1 Introduction

Let us consider the following simple example of ordinary differential equation

$$u'(x) = \frac{\sin(x)}{x},\ x \in (1, 2), \qquad [1]$$

with

$$u(1) = 1. \qquad [2]$$

It is well known, that the solution of the ordinary differential equation [1]–[2] is

$$u(x) = 1 + \int_1^x \frac{\sin(t)}{t} dt, \;\; x \in (1, 2). \qquad [3]$$

Since we do not know the exact value of the integral $\int_1^x \frac{\sin(t)}{t} dt$, one does not know exactly the expression of $u(x)$ defined by [3], for all $x \in (1, 2)$. We think then about the following options in order to compute approximatly $u(x)$:

- We approximate the integral $\int_1^x \frac{\sin(t)}{t} dt$ using methods of numerical integration
- We use the known numerical methods to approximate equation [1]–[2]. The advantage of this last option is that we do not need an expression for u, like that of [3], and then we approximate directly equation [1]–[2]. Among the numerical methods which allow us to approximate [1]–[2], we have:
  - Finite difference methods
  - Finite element methods
  - Finite volume methods

2 A simple example and some questions to be asked in finite difference methods

In order to justify the convergence of a finite difference method approximating a differential equation, let us consider the following simple equation: find $u \in \mathcal{C}^1(0, 1)$ such that :

$$u'(x) = 2x, \;\; x \in (0, 1), \qquad [4]$$

with

$$u(0) = 0. \qquad [5]$$

The solution of the previous equation is:

$$u(x) = x^2, \;\; x \in (0, 1). \qquad [6]$$

Let us consider a positive parameter h which is expected to goes to zero, and consider the points $0 = x_0 < x_1 < \ldots x_N = 1$ such that $x_i - x_{i-1} = h$, for all $i \in \{1, \ldots, N\}$. This yields the following explicit expression:

$$x_i = ih, \;\; \forall i \in \{0, \ldots, N\}. \qquad [7]$$

It is useful to relate h and N; indeed $Nh = 1$ implies that

$$h = \frac{1}{N}. \tag{[8]}$$

The aim now is to compute the value of u on x_i, for all $i \in \{1, \dots, N\}$ (Recall that for $i = 0$, $u(x_i) = u(x_0) = u(0) = 0$.). To do so, we replace x in [9] by x_i to get

$$u'(x_i) = 2x_i, \ \forall i \in \{0, \dots, N\}. \tag{[9]}$$

Using a simple formula of Taylor's expansion, we get, for some $\xi_i \in (x_i, x_{i+1})$

$$\frac{u(x_{i+1}) - u(x_i)}{h} = u'(x_i) + \frac{h}{2}u''(\xi_i), \ \forall i \in \{0, \dots, N-1\}. \tag{[10]}$$

Which gives

$$u'(x_i) = \frac{u(x_{i+1}) - u(x_i)}{h} - \frac{h}{2}u''(\xi_i), \ \forall i \in \{0, \dots, N-1\}. \tag{[11]}$$

Inserting this in [9], we get

$$\frac{u(x_{i+1}) - u(x_i)}{h} - \frac{h}{2}u''(\xi_i) = 2x_i, \ \forall i \in \{0, \dots, N-1\}. \tag{[12]}$$

Which yields

$$\frac{u(x_{i+1}) - u(x_i)}{h} = 2x_i + \frac{h}{2}u''(\xi_i), \ \forall i \in \{0, \dots, N-1\}. \tag{[13]}$$

Since we know already that $u'(x) = x$, for all $x \in (0,1)$, then $u''(x) = 2$, for all $x \in (0,1)$. But even we know this, we neglect the second term on the right hand side of [13] because of the fact that we assume that h is "small", and we denote by u_i an approximation to $u(x_i)$. Therefore expansion [13] becomes as

$$\frac{u_{i+1} - u_i}{h} = 2x_i, \ \forall i \in \{0, \dots, N-1\}, \tag{[14]}$$

where, since $u(0) = 0$, it is convenient to set, since u_0 is expected to approximate $u(0)$,

$$u_0 = 0. \tag{[15]}$$

This implies that

$$u_{i+1} = u_i + 2x_i h, \ \forall i \in \{0, \dots, N-1\}. \tag{[16]}$$

This implies

$$u_{i+1} = u_0 + 2h \sum_{i=0}^{i} x_j, \ \forall i \in \{0, \dots, N-1\}. \tag{[17]}$$

Replacing $i+1$ by i in [17], we get

$$u_i = u_0 + 2h \sum_{i=0}^{i-1} x_j, \ \forall i \in \{1, \dots, N\}. \tag{[18]}$$

therefore the expression of u_i, given in [18], could be written as

$$u_i = 2h \sum_{i=0}^{i-1} x_j, \ \forall i \in \{1, \dots, N\}. \tag{[19]}$$

Therefore, thanks to $x_j = jh$ and using the fact that $\sum_{i=0}^{i-1} j = \frac{(i-1)i}{2}$., expression [19] becomes as

$$
\begin{aligned}
u_i &= u_0 + 2h\sum_{i=0}^{i-1} x_j \\
&= 0 + 2h\sum_{i=0}^{i-1} jh \\
&= 2h^2\sum_{i=0}^{i-1} j \\
&= 2h^2\frac{(i-1)i}{2} \\
&= h^2(i-1)i \\
&= (h(i-1))\,(ih) \\
&= x_i x_{i-1}
\end{aligned}
\qquad [20]
$$

Let us denote by u_h the vector $(u_i)_0^N$. The question now: is u_h converges to u, as $h \to 0$, in the following sense for example:

$$\max_{i=0}^{N} |u(x_i) - u_i| \to 0, \text{ as } h \to 0? \qquad [21]$$

We have, since $u(x_i) = x_i^2$, for all $i \in \{0, \dots, N\}$, using the expression of u_i given by [19] and $x_i \le 1$

$$
\begin{aligned}
|u(x_i) - u_i| &= |x_i^2 - x_i x_{i-1}| \\
&= x_i |x_i - x_{i-1}| \\
&= x_i h \\
&\le h.
\end{aligned}
\qquad [22]
$$

When $h \to 0$ in the previous inequality, we get

$$|u(x_i) - u_i| \to 0, \text{ as } h \to 0. \qquad [23]$$

So far, we have proven the convergence of the finite difference approximate solution $u_h = (u_i)_0^N$, given by [19] and [15], towards the exact solution u thanks to the explicite expression of u given by [6]. Let us now prove this convergence without use of the expression of [6] of u.

Remark 1 (Finite difference solution through matrix) The problem ([14],[15]) could be written as:

$$\mathcal{A}u_h = f_h, \qquad [24]$$

where $\mathcal{A}$ is a matrix of order $N-1$ and $u_h = (u_1, u_2, \dots, u_{N-1})^t$ is the vector whose de components are the unknowns $u_1, u_2, \dots, u_{N-1})^t$ defined by ([14],[15]), with $u_0 = 0$, and $f_h = (2x_1, 2x_2, \dots, 2x_{N-1})^t$. Therefore, according to ([14],[15]), the i-th component of $\mathcal{A}u_h$ is $\frac{u_{i+1} - u_i}{h}$.

2.1 A convergence proof of the finite difference solution [19] and [15] without make appeal to [6]

Substracting [14] from [13], we get

$$\frac{e_{i+1}-e_i}{h} = \frac{h}{2}u''(\xi_i),\ \forall i \in \{0,\ldots,N-1\}, \tag{25}$$

where $e_i = u(x_i) - u_i$, for all $i \in \{0,...,N\}$.

Mutiplying both sides of [25] by h and adding e_i to the both sides of the result, we get

$$e_{i+1} = e_i + \alpha_i,\ \forall i \in \{0,\ldots,N-1\}. \tag{26}$$

Let us denote by α_i to the value $h\frac{h}{2}u''(\xi_i)$, for all $i \in \{0,\ldots,N-1\}$

As done before

$$e_i = e_0 + \sum_{j=0}^{i-1}\alpha_j,\ \forall i \in \{1,\ldots,N\}. \tag{27}$$

Since $e_0 = u(x_0) - u_0 = u(0) - u_0 = 0$, then

$$e_i = \sum_{j=0}^{i-1}\alpha_j,\ \forall i \in \{1,\ldots,N\}. \tag{28}$$

Let us assume the following assumption on u, there exists a positive constant M such that

$$|u''(x)| \le M,\ \forall x \in [0,1]. \tag{29}$$

(This assumption could be deduced, from instance, from equation by differentiating [4], and then $u''(x) = 2$ for all $x \in (0,1)$. Such assumptions on the derivatives of the exact solution, like that of [29], are used mainly when we need to prove the convergence or to determine the *convergence order*, see next sections.)

Estimate [29] with the fact that $alpha_i = h\dfrac{h}{2}u''(\xi_i)$ implies that

$$|\alpha_i| \le Mh^2,\ \forall i \in \{0,\ldots,N-1\}, \tag{30}$$

which implies that, using [28] and [8]

$$\begin{aligned} |e_i| &\le Mh^2\sum_{j=0}^{i-1}1 \\ &\le MNh^2 \\ &= Mh,\ \forall i \in \{1,\ldots,N\} \end{aligned} \tag{31}$$

From this simple example, we deduce the basic concepts of the finite difference methods.

2.2 Basic concepts of finite difference methods

- finite difference method is a method aims to approximate differential and partial differential equations

- finite difference method allows us to approximate the exact solution on some points. These points are called *mesh points*.
- finite difference method is based on the approximation of the derivatives which appear in differential or partial differential equation using Taylor expansion.

3 A second example

For more understanding to how to apply the previous steps of finite difference discretization, let us consider the following example

$$u'(x) - \alpha u(x) = 0, \ x \in (0,1), \tag{32}$$

with the following "boundary conditions":

$$u(0) = 1, \tag{33}$$

where α is some given real number.
The solution of [32]–[33] is

$$u(x) = \exp(\alpha x). \tag{34}$$

Let us move now to descretize problem [32]–[33] by finite difference methods. To this end, we consider a mesh step h, and the mesh points $x_i = ih$, for all $i \in \{0, \ldots, N\}$, where $x_0 = 0$ qnd $x_N = 1$. Therefore $Nh = 1$.
Replacing x by x_i in [32], we get, for u "smooth enough"

$$u'(x_i) - \alpha u(x_i) = 0, \ \forall i \in \{0, \ldots, N\} \tag{35}$$

Let us approximate $u'(x_i)$ by $\frac{u(x_{i+1} - u(x_i)}{h}$, and denote by u_i an approximation to $u(x_i)$. Therefore, u_i satisfies the following problem

$$\frac{u_{i+1} - u_i}{h} - \alpha u_i = 0, \ \forall i \in \{0, \ldots, N-1\}, \tag{36}$$

and

$$u_0 = 1. \tag{37}$$

Multiplying both sides of [36] by h, and adding $u_i + \alpha u_i$ to the both sides of the result, we get

$$u_{i+1} = (1 + \alpha h)\, u_i, \ \forall i \in \{0, \ldots, N-1\}, \tag{38}$$

which gives

$$u_i = (1 + \alpha h)^i, \ \forall i \in \{0, \ldots, N\}. \tag{39}$$

Let us move now to justify the convergence of $u_h = (u_i)_1^N$ towards the solution u in the sense that $\max_{i \in \{1, \ldots, N\}} |u(x_i) - u_i| \to 0$.
Indeed, let us assume that the expression u defined by [34] is known. In case when the expression

of the exact solution is not known, which is the general case of the equations to be solved, we need to perform some techniques based on the equation satisfied by the exact solution u, see below. Indeed, using a Taylor's expansion, we get, since $x_i = ih$

$$\begin{aligned} u_i &= e^{i\left(\alpha h - \frac{\alpha^2 h^2}{2} + \alpha^2 h^2 \varepsilon_1(h)\right)} \\ &= e^{i\alpha h} e^{-x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right)} \\ &= e^{\alpha x_i} \left\{1 - x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right) + x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right) \varepsilon_2 \left(-x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right)\right)\right\} \\ &= u(x_i) + \mathcal{A}_h, \end{aligned} \tag{40}$$

where

$$\mathcal{A}_h = -x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right) e^{\alpha x_i} \left\{1 - \varepsilon_2 \left(-x_i h \alpha^2 \left(\frac{1}{2} - \varepsilon_1(h)\right)\right)\right\}. \tag{41}$$

We have used the following Taylor expansions:

$$\log(1 + x) = x - \frac{x^2}{2} + x^2 \varepsilon(x), \tag{42}$$

and

$$e^x = 1 + x + x\varepsilon_2(x), \tag{43}$$

such that

$$\varepsilon_1(x) \to 0, \text{ and } \varepsilon_2(x) \to 0, \text{ as } x \to 0. \tag{44}$$

Since $\varepsilon_1(h) \to 0$, as $h \to 0$ then, for sufficiently small h, there exists a positivie number C_1 such that

$$|\varepsilon_1(h)| \le C_1. \tag{45}$$

On the other hand, since $x_i \in [0, 1]$, then $-x_i h \alpha^2 \left(\frac{1}{2} + \varepsilon_1(h)\right) \to 0$, as $h \to 0$. This last limit combined with the fact that $\varepsilon_2(x) \to 0$ as $x \to 0$ implies that $\varepsilon_2 \left(-x_i h \alpha^2 \left(\frac{1}{2} + \varepsilon_1(h)\right)\right) \to 0$, as $h \to 0$. Therefore, for sufficiently small h, there exists a positivie number C_2 such that

$$|\varepsilon_2 \left(-x_i h \alpha^2 \left(\frac{1}{2} + \varepsilon_1(h)\right)\right)| \le C_2. \tag{46}$$

This with [45], [46], and the fact that $x_i \in [0, 1]$, implies that, for a sufficiently small h,

$$|\mathcal{A}_h| \le C_3 h, \tag{47}$$

where

$$C_3 = \alpha^2 \left(\frac{1}{2} + C_1\right)(1 + C_2)e^{\alpha}. \tag{48}$$

This with [40] implies that, for a sufficiently small h

$$|u_i - u(x_i)| \le C_3 h, \ \forall i \in \{1, \ldots, N\} \tag{49}$$

Remark 2 (Finite difference solution through matrix) The problem [36]–[37] could be written as:

$$\mathcal{A}u_h = 0, \tag{50}$$

where $\mathcal{A}$ is a matrix order $N-1$ and $u_h = (u_1, u_2, \ldots, u_{N-1})^t$ is the vector whose de components are the unknowns $u_1, u_2, \ldots, u_{N-1})^t$ defined by [36]–[37], with $u_0 = 0$. Therefore, according to [36]–[37], the i-th component of $\mathcal{A}u_h$ is $\dfrac{u_{i+1}-u_i}{h} - \alpha u_i$.

3.1 A convergence proof without make appeal to the expression [34] of u

We proceed as in [2.1] to the prove the of the finite difference approximate solution [39] towards the exact solution of [32]–[33] without make appeal to the expression [34] of u.

Inserting Taylor's expansion [10] in Equation

$$\frac{u(x_{i+1})-u(x_i)}{h} - \frac{h}{2}u''(\xi_i) - \alpha u(x_i) = 0, \ \forall i \in \{0,\ldots,N-1\}. \tag{51}$$

Adding $\dfrac{h}{2}u''(\xi_i)$ to the both sides of the resulting equation, we get

$$\frac{u(x_{i+1})-u(x_i)}{h} - \alpha u(x_i) = \frac{h}{2}u''(\xi_i). \ \forall i \in \{0,\ldots,N-1\}. \tag{52}$$

Substracting [36] from [52], we get

$$\frac{e_{i+1}-e_i}{h} - \alpha e_i = \frac{h}{2}u''(\xi_i). \ \forall i \in \{0,\ldots,N-1\}, \tag{53}$$

where $e_i = u(x_i) - u_i$, for all $i \in \{0,\ldots,N\}$.

Multiplying both sides of [53] by h and adding $(1+h\alpha)e_i$ to the both sides of the result, we get

$$e_{i+1} = (1+h\alpha)e_i + \alpha_i, \ \forall i \in \{0,\ldots,N-1\}, \tag{54}$$

where α_i is defined by $h\frac{h}{2}u''(\xi_i)$.

Relation [54]

$$e_i = (1+h\alpha)^i e_0 + \sum_{j=0}^{i-1}(1+h\alpha)^{i-j-1}\alpha_j, \ \forall i \in \{1,\ldots,N\}. \tag{55}$$

Since $e_0 = 0$, then the expression [55] becomes as

$$e_i = \sum_{j=0}^{i-1}(1+h\alpha)^{i-j-1}\alpha_j, \ \forall i \in \{1,\ldots,N\}. \tag{56}$$

Let us estimate e_i using previous expression. Indeed, since $1+|h\alpha \geq 1$ and thanks to [30], using the fact that $\log(1+x) \leq x$ for all $x \geq 0$, we have

$$
\begin{aligned}
|e_i| &\leq \sum_{j=0}^{i-1} |1+h\alpha|^{i-j-1}\alpha_j \\
&\leq \sum_{j=0}^{i-1} |1+h\alpha|^{i-j-1}|\alpha_j| \\
&\leq \sum_{j=0}^{i-1} (1+h|\alpha|)^{i-j-1}|\alpha_j| \\
&\leq h(1+h|\alpha|)^N M \\
&= hM(1+h|\alpha|)^{\frac{1}{h}} \\
&= hMe^{\frac{\log(1+h|\alpha|)}{h}} \\
&= hMe^{\frac{\log(1+h|\alpha|)}{h}} \\
&\leq hMe^{|\alpha|}.
\end{aligned} \quad [57]
$$

4 A third example

In the previous, we considered two examples in which the finite difference approximate solution is defined explicitly in the sense we could compute the unknowns of the discrete problem explicitly. In this Subsection, we consider an example in which the unkowns of the discrete problem are not computed explicitly; more precise the finite difference approximate solution is a solution of a system. Let us consider the following differential equation:

$$-u''(x) = \pi^2 \sin(\pi x),\ x \in (0,1), \quad [58]$$

with, say Dirichlet boundary conditions

$$u(0) = u(1) = 0. \quad [59]$$

To this end, we consider a mesh step h, and the mesh points $x_i = ih$, for all $i \in \{0, \ldots, N\}$, where $x_0 = 0$ qnd $x_N = 1$. Therefore $Nh = 1$.

Replacing x by x_i in [58], we get, for u "smooth enough"

$$-u''(x_i) = \pi^2 \sin(\pi x_i),\ i \in \{0, \ldots, N\}. \quad [60]$$

We have, thanks to Taylor's expansion, for some $\xi_i \in (x_i, x_{i+1})$

$$\frac{u(x_{i+1}) - u(x_i)}{h} = u'(x_i) + \frac{h}{2}u''(x_i) + \frac{h^2}{6}u^{(3)}(x_i) + \frac{h^3}{24}u^{(4)}(\xi_i),\ \forall i \in \{0, \ldots, N-1\} \quad [61]$$

and, for $\bar{\xi}_i \in (x_i, x_{i+1})$then

$$\frac{u(x_i) - u(x_{i-1})}{h} = u'(x_i) - \frac{h}{2}u''(x_i) + \frac{h^2}{6}u^{(3)}(x_i) - \frac{h^3}{24}u^{(4)}(\bar{\xi}_i),\ \forall i \in \{1, \ldots, N\} \quad [62]$$

Substracting [62] from [61], we get

$$\frac{u(x_{i+1} - 2u(x_i) + u(x_{i-1})}{h} = hu''(x_i) + \frac{h^3}{24}(u^{(4)}(\xi_i) + u^{(4)}(\bar{\xi}_i)),\ \forall i \in \{1, \ldots, N-1\}. \quad [63]$$

Dividing previous equality by h, we get

$$\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = u''(x_i) + \beta_i,\ \forall i \in \{1, \ldots, N-1\}, \quad [64]$$

where

$$\beta_i = \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\bar{\xi}_i)). \quad [65]$$

Thanks to [60], we get $u''(x_i) = -\pi^2 \sin(\pi x_i)$ for all $i \in \{1, \ldots, N-1\}$; inserting this in equality [64] to get

$$\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = -\pi^2 \sin(x_i) + \beta_i,\ \forall i \in \{1, \ldots, N-1\}. \quad [66]$$

by neglecting the term β_i

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = \pi^2 \sin(\pi x_i),\ \forall i \in \{1, \ldots, N-1\}, \quad [67]$$

where u_i is an approximation of $u(x_i)$, for all $i \in \{0, \ldots, N\}$. Since $u(0) = u(1) = 0$, we chose

$$u_0 = u_N = 0. \quad [68]$$

Let $u_h = (u_i)_0^N$ be defined by [67]–[68].

4.1 How to compute the finite difference approximate solution u_h defined by [67]–[68]

To compute finite difference approximate solution u_h defined by [67]–[68], we two possiblities, either

- we have to resolve an algebraic system:

$$\mathcal{A}u_h = f_h, \quad [69]$$

  where $(\mathcal{A}u_h)_i = -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$ and $(f_h)_i = (\pi^2 \sin(\pi x_i))$. We justify now that there exists a unique u_h satisfying [69], or
- we compute exlicitly u_i

We study now each possiblity

4.1.1 We resolve the system [69]

Let us justify the existence and uniqueness of the solution of [69]. To do so, one remarks that $\mathcal{A}$ is a square matrix, one could deduce that $\mathcal{A}$ is injective yields the sujectivity of $\mathcal{A}$. This means that the uniqueness of the solution of [69] yields the existence of the solution of [69].

It suffices then to justify that there exists at most one solution u_h for [69]. We assume that there exists a vector $\omega_h = (\omega_i)_1^N$ such that

$$\mathcal{A}\omega_h = 0, \quad [70]$$

and

$$\omega_0 = \omega_N = 0. \qquad [71]$$

Therefore, using the definition of the matrix $\mathcal{A}$ to get

$$\omega_{i+1} - \omega_i = \omega_i - \omega_{i-1},\ \forall i \in \{1, \ldots, N-1\} \qquad [72]$$

Summing [73] over $i \in \{1, \ldots, j\}$ to get, since $\omega_0 = 0$

$$\omega_{j+1} - \omega_1 = \omega_j,\ \forall j \in \{1, \ldots, N-1\}, \qquad [73]$$

which gives

$$\omega_{j+1} - \omega_j = \omega_1,\ \forall j \in \{1, \ldots, N-1\}, \qquad [74]$$

Summing [74] over $j \in \{1, \ldots, N-1\}$ to get, since $\omega_N = 0$

$$-\omega_1 = (N-1)\omega_1, \qquad [75]$$

which implies that

$$\omega_1 = 0. \qquad [76]$$

This with [73] implies that

$$\omega_{j+1} = \omega_j,\ \forall j \in \{1, \ldots, N-1\}, \qquad [77]$$

which yields

$$\omega_j = 0,\ \forall j \in \{2, \ldots, N\}. \qquad [78]$$

This with [71] yields

$$\omega_j = 0,\ \forall j \in \{0, \ldots, N\}. \qquad [79]$$

4.1.2 We compute u_i

In the previous subsection, we used the matrix form [69], to prove the existence and uniqueness of the solution of [67]–[68]. It is also possible to compute explicitly the solution u_i of [67]–[68]. The the advantage of the use of the matrix form [69] to prove the existence and uniqueness is that it is more general and we do not need to compute explicitly u_i.

To compute u_i, we multiply equality [67] by h^2 to get

$$u_{i+1} - u_i - (u_i - u_{i-1}) = -h^2\pi^2 \sin(\pi x_i),\ \forall i \in \{1, \ldots, N-1\}, \qquad [80]$$

Summing over $i \in \{1, j-1\}$, for $j \in \{1, N-1\}$, and using the fact that $u_0 = 0$ to get

$$u_{j+1} - u_j - u_1 = -h^2 \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i),\ \forall j \in \{1, \ldots, N-1\}. \qquad [81]$$

Summing previous equality on $j \in \{1, N-1\}$ and using the fact $u_N = 0$ to get

$$-Nu_1 = -h^2 \sum_{j=1}^{N-1} \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i),\ \forall i \in \{1, \ldots, N-1\}. \qquad [82]$$

Which implies that, since $N = 1/h$

$$u_1 = h^3 \sum_{j=1}^{N-1} \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i). \qquad [83]$$

After having computed u_1, let us compute u_i for all $i \in \{2, \ldots, N-1\}$. Summing [81] over $j \in \{1, \ldots, k-1\}$ to get

$$u_k - k u_1 = -h^2 \sum_{j=1}^{k-1} \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i),\ \forall k \in \{2, \ldots, N-1\}, \qquad [84]$$

which implies, using [83]

$$u_k = k h^3 \sum_{j=1}^{N-1} \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i) - h^2 \sum_{j=1}^{k-1} \sum_{i=1}^{j-1} \pi^2 \sin(\pi x_i),\ \forall k \in \{2, \ldots, N-1\}. \qquad [85]$$

4.1.3 The *convergence order* of the finite difference solution [67]–[68]

Let $e_i = u(x_i) - u_i$ for all $i \in \{0, \ldots, N\}$. Subtracting [67] from [66] to get

$$-\frac{e_{i+1} - 2e_i + e_{i-1}}{h^2} = \beta_i,\ \forall i \in \{1, \ldots, N-1\}, \qquad [86]$$

where

$$e_0 = e_N = 0. \qquad [87]$$

Using the same reasoning of the previous subsection, we get

$$e_k = k h^3 \sum_{j=1}^{N-1} \sum_{i=1}^{j-1} \beta_i - h^2 \sum_{j=1}^{k-1} \sum_{i=1}^{j-1} \beta_i,\ \forall k \in \{2, \ldots, N-1\}. \qquad [88]$$

Let us assume that there exists a positive constant M such that

$$|u^{(4)}(x)| \le M,\ \forall x \in (0,1), \qquad [89]$$

therefore the following estimate for β_i holds

$$|\beta_i| \le \frac{M}{12} h^2. \qquad [90]$$

Using this in [88] to get

$$|e_k| \le k h^3 \frac{M}{12} h^2 N^2 + h^2 \frac{M}{12} h^2 N^2,\ \forall k \in \{2, \ldots, N-1\}. \qquad [91]$$

which yields since $k \le N$ and $Nh = 1$

$$|e_k| \le h^2 \frac{M}{6},\ \forall k \in \{2, \ldots, N-1\}. \qquad [92]$$

Remark 3 (An approximation of order h^2 to $\frac{u(x_{i+1})-u(x_i)}{h}$) Estimate [92] implies that u_i approximates $u(x_i)$ by order h^2. Some times, we do not only need to approximate $u(x_i)$ but also we need to approximate $u'(x_i)$. We can use the estimate [92] to prove that $\frac{u_{i+1}-u_i}{h}$ approximate $u'(x_i)$ by order h. Indeed,using the triangle inequality, estimate [92] to get (recall that $e_i = u(x_i) - u_i$)

$$\begin{aligned}
|\frac{u_{i+1}-u_i}{h} - u'(x_i)| &\leq |\frac{u_{i+1}-u_i}{h} - \frac{u(x_{i+1})-u(x_i)}{h}| + |\frac{u(x_{i+1})-u(x_i)}{h} - u'(x_i)| \\
&\leq |\frac{u_{i+1}-u_i}{h} - \frac{u(x_{i+1})-u(x_i)}{h}| + h \max_{x\in[0,1]} |u''(x)| \\
&\leq \frac{1}{h}\{|e_{i+1}| + |e_i|\} + h \max_{x\in[0,1]} |u''(x)| \\
&\leq h\frac{M}{3} + h \max_{x\in[0,1]} |u''(x)| \\
&\leq (\frac{M}{3} + \max_{x\in[0,1]} |u''(x)|)h \qquad [93]
\end{aligned}$$

But we can prove that $\dfrac{u_{i+1}-u_i}{h}$ approximate $\dfrac{u(x_{i+1})-u(x_i)}{h}$ by order h^2 in some discrete L^2–norm. Indeed, [86] implies

$$-\frac{e_{i+1}-e_i}{h} + \frac{e_i - e_{i-1}}{h} = h\beta_i e_i, \ \forall i \in \{1, \ldots, N-1\}. \qquad [94]$$

Multiplying both sides of [94] by e_i, summing over $i \in \{1, N-1\}$, reordering the sum in the left hand side, and using [87], we get

$$\sum_0^{N-1} h(\frac{e_{i+1}-e_i}{h})^2 = \sum_{1,N-1} h\beta_i e_i. \qquad [95]$$

Since $|\sum_1^{N-1} h\beta_i| \leq h^2\dfrac{M}{6}$, then [95] yields

$$\sum_0^{N-1} h(\frac{e_{i+1}-e_i}{h})^2 \leq h^2\frac{M}{6}\sum_1^{N-1} h|e_i|. \qquad [96]$$

Using now the following discrete version of Poincaré inequality, for some positive constant independent of h

$$\sum_1^{N-1} h|e_i| \leq C\sum_0^{N-1} h(\frac{e_{i+1}-e_i}{h})^2. \qquad [97]$$

This with [96] implies that

$$\left(\sum_0^{N-1} h(\frac{e_{i+1}-e_i}{h})^2\right)^{\frac{1}{2}} \leq h^2\frac{M}{6}. \qquad [98]$$

5 What we need to approximate a differential equation by finite difference method

From the previous examples, we can guess which material we need for finite difference method. The following Subsections are dealt with this material.

5.1 Taylor expansions

Let $n \in \mathbb{N}^\star$, and a and b be two real numbers. Let f be a sufficiently smooth function on an interval (a, b), namely $f \in \mathcal{C}^n(a, b)$. Let $x_0 \in (a, b)$, for any $h \in \mathbb{R}$ such that $x_0 + h \in (a, b)$, there exists a function $\varepsilon(h)$ such that

$$f(x_0 + h) = f(x_0) + \frac{f'(x_0)}{1!} + h^2\frac{f''(x_0)}{2!} + h^3\frac{f^{(3)}(x_0)}{3!} + \ldots + h^n\frac{f^{(n)}(x_0)}{n!} + h^n\varepsilon(h), \qquad [99]$$

and

$$\varepsilon(h) \to 0, \text{ as } h \to 0. \qquad [100]$$

5.2 Consideration of mesh

Assume that we have to approximate a differential equation posed on interval $\mathcal{I}$. Recall that the aim of finite difference method is to approximate the exact solution on some points belong to $\mathcal{I}$. These points called *mesh points.*

5.3 Computing the finite difference approximate solution

After having approximated the derivatives which appear in the diferential equation to be solved, by using Talyor expansions, we replace the variable x by xi and then we obtain a finite difference approximate solution, denoted by uh.This finite difference solution uh is defined either by:

- an explicite expression for the finite difference approximate solution: this means that we can compute ui explicitly, or
- by an algebraic system to be solved

6 How to prove the convergence of a finite difference approximate solution

6.1 Introduction: some concepts

In the previous examples,we have proven the convergence of the finite difference solution using two methods:

- First method: we use an explicit expression for the exact solution as well as an explicit expresion for the finite difference solution, and then we make the difference between these two expressions.
- second method: let us assume that the finite difference solution is defined as the solution of a problem could be written as:

$$\mathcal{L}_h u_h = f_h \qquad [101]$$

where u_h is the finite difference solution, $\mathcal{L}_h$ is an operator and may be is non–linear (note that $\mathcal{L}_h$ is a matrix in all the examples we treated before).
Since the finite difference solution u_h is some vector in which its components are expected to approximate the values of the exact solution on the mesh points, it is possible then to act $\mathcal{L}_h$ on u by considering u as a vector, denoted by $[u]_h$, in which each component of $[u]_h$ is the value of u on the mesh point to which the corresponding component of u_h is expcted to approximate. Let us assume that, we get, usually this could be obtained thanks to Taylor's expansions

$$\mathcal{L}_h[u]_h = f_h + \varepsilon_h \qquad [102]$$

To prove the convergence of u_h towards u, we assume that

- the "remainder term" ε_h satisfies, for some norm denoted by $\|\cdot\|_{\mathcal{F}}$, the following convergence holds:

$$\|\varepsilon_h\|_{\mathcal{F}} \to 0, \text{ as } h \to 0. \qquad [103]$$

- the operator $\mathcal{L}_h$ is invertible and the following continuity of $\mathcal{L}_h^{-1}$ holds, for some constant C independent of the mesh parameter h:

$$\|\mathcal{L}_h^{-1} f_h\|_{\mathcal{U}} \le C\|f_h\|_{\mathcal{F}} \qquad [104]$$

Substracting [101] from [102], we get

$$\mathcal{L}_h([u]_h - u_h) = \varepsilon_h. \qquad [105]$$

Which gives

$$u - u_h = \mathcal{L}_h^{-1}(\varepsilon_h). \qquad [106]$$

This implies that, with [104]

$$\|u - u_h\|_{\mathcal{U}} \le C\|\varepsilon_h\|_{\mathcal{F}}, \qquad [107]$$

which yields

$$\|u - u_h\|_{\mathcal{U}} \to 0, \text{ as } h \to 0. \qquad [108]$$

6.2 Some simple examples

Since the first method in the previous Subsection, that is the convergence proof through the computation of the exact unknown solution and the finite difference solution, can not be applied in the general case, we will devote this Subsection to provide with some examples in which we explain how to apply the concepts stated in second method of the previous Subsection. We will not only apply the concepts of Subsection 6.1 on the examples treated in Sections 2, 3, and 4, but also we apply these concepts on other examples in which the concepts of Subsection 6.1 are not obvious to apply on.

- First example In this item, we apply the concepts stated in the second method of Subsection 6.1, on the example of Section 2, that is the finite difference approximation [14]–[15] of [4]–[5]:
  - property [103]: let us first set [14]–[15] in the form of [101]. Indeed, $\mathcal{L}_h$ is a square matrix with N lines, and

$$\mathcal{L}_h u_h = \left(\frac{u_1 - u_0}{h}, \dots, \frac{u_N - u_{N-1}}{h}\right)^t, \qquad [109]$$

where $u_0 = 0$, $u_h = (u_1, u_2, \dots, u_N)^t$, and

$$f_h = (2x_0, \dots, 2x_{N-1})^t. \qquad [110]$$

By acting the matrix $\mathcal{L}_h$ on the function u with replacing u_i by $u(x_i)$, for all $i \in \{0, \dots, N\}$, we get

$$\mathcal{L}_h u = \left(\frac{u(x_1) - u(x_0)}{h}, \dots, \frac{u(x_N) - u(x_{N-1})}{h}\right)^t, \qquad [111]$$

Using the Taylor exapansion [13], we get

$$\begin{aligned} \mathcal{L}_h u &= \left(\frac{u(x_1) - u(x_0)}{h}, \dots, \frac{u(x_N) - u(x_{N-1})}{h}\right)^t \\ &= \left(2x_0 + \frac{h}{2}u''(\xi_0), \dots, 2x_{N-1} + \frac{h}{2}u''(\xi_{N-1})\right)^t \\ &= (2x_0, \dots, 2x_{N-1})^t + \left(\frac{h}{2}u''(\xi_0), \dots, \frac{h}{2}u''(\xi_{N-1})\right)^t \\ &= (2x_0, \dots, 2x_{N-1})^t + \frac{h}{2}\left(u''(\xi_0), \dots, u''(\xi_{N-1})\right)^t \\ &= f_h + \frac{h}{2}\left(u''(\xi_0), \dots, u''(\xi_{N-1})\right)^t. \end{aligned} \qquad [112]$$

Now the function ε_h given by [102] is defined by

$$\varepsilon_h = \frac{h}{2}\left(u''(\xi_0), \dots, u''(\xi_{N-1})\right)^t. \qquad [113]$$

By assuming assumption [29], we get since $\frac{1}{2} < 1$

$$\|\varepsilon_h\|_\infty \le Mh, \qquad [114]$$

where $\|\cdot\|_\infty$ denotes the uniform–norm

$$\|(s_0, \dots, s_{N-1})\|_\infty = \max_{i=0}^{N-1}(|s_0|, \dots, |s_{N-1}|). \qquad [115]$$

In particular, estimate [114] implies the convergence

$$\|\varepsilon_h\|_\infty \to 0, \text{ as } h \to 0. \qquad [116]$$

  - property [104]: it suffices to prove that if $\mathcal{L}_h u_h = f_h$, where $u_h = (u_1, \dots, u_N)^t$ and $f_h = (f_0, f_1, \dots, f_{N-1})^t$, we have the following estimate, for some positive constant independent of the parameter h

$$\|u_h\|_\infty \le C\|f_h\|_\infty. \qquad [117]$$

this implies

* the matrix $\mathcal{L}_h$ is injective, since $f_h = 0$ in [117] implies $u_h = 0$,
* since $\mathcal{L}_h$ is a square matrix, the previous injectivity of $\mathcal{L}_h$ implies the surjectivity,
* estimate [117] yields [117].

Using the computations [14]–[19] combined with the triangle inequality and the fact that $u_0 = 0$, we get

$$|u_i| \le h \sum_{j=0}^{i-1} |f_j|, \ \forall i \in \{1, \dots, N\}. \tag{118}$$

Which implies, since $i \le N$

$$|u_i| \le hN \|f_h\|_\infty, \ \forall i \in \{1, \dots, N\}. \tag{119}$$

Therefore, since $Nh = 1$

$$\|u_h\|_\infty \le \|f_h\|_\infty, \tag{120}$$

which means that [117] holds for all $1 \le C$.

- Second example In this item, we apply the concepts stated in the second method of Subsection 6.1, on the example of Section 3, that is the finite difference approximation [36]–[37] of [32]–[33]:

  – property [103]: let us first set [36]–[37] in the form of [101]. Indeed, $\mathcal{L}_h$ is a square matrix with N lines, and

$$\mathcal{L}_h u_h = \left(\frac{u_1 - u_0}{h} - \alpha u_0, \dots, \frac{u_N - u_{N-1}}{h} - \alpha u_{N-1} \right)^t, \tag{121}$$

where $u_0 = 0$, $u_h = (u_1, u_2, \dots, u_N)^t$, and f_h is the vector of N components

$$f_h = (0, \dots, 0)^t. \tag{122}$$

By acting the matrix $\mathcal{L}_h$ on the function u with replacing u_i by $u(x_i)$, for all $i \in \{0, \dots, N\}$, we get

$$\mathcal{L}_h u = \left(\frac{u(x_1) - u(x_0)}{h} - \alpha u(x_0), \dots, \frac{u(x_N) - u(x_{N-1})}{h} - \alpha u(x_{N-1}) \right)^t, \tag{123}$$

Using the Taylor exapansion [52], we get

$$\begin{aligned} \mathcal{L}_h u &= \left(\frac{u(x_1) - u(x_0)}{h} - \alpha u(x_0), \dots, \frac{u(x_N) - u(x_{N-1})}{h} - \alpha u(x_{N-1}) \right)^t \\ &= \left(\frac{h}{2} u''(\xi_0), \dots, \frac{h}{2} u''(\xi_{N-1}) \right)^t \\ &= f_h + \frac{h}{2} \left(u''(\xi_0), \dots, u''(\xi_{N-1}) \right)^t. \end{aligned} \tag{124}$$

Now the function ε_h given by [102] is defined by

$$\varepsilon_h = \frac{h}{2} \left(u''(\xi_0), \dots, u''(\xi_{N-1}) \right)^t. \tag{125}$$

With the assumption the second derivative of u is bounded uniformly by a positive constant M, we get since $\frac{1}{2} < 1$

$$\|\varepsilon_h\|_\infty \leq Mh, \tag{126}$$

In particular, estimate [126] implies the convergence

$$\|\varepsilon_h\|_\infty \to 0, \text{ as } h \to 0. \tag{127}$$

– property [104]: it suffices to prove that if $\mathcal{L}_h u_h = f_h$, where $u_h = (u_1, \ldots, u_N)^t$ and $f_h = (0, \ldots, 0)^t$, we have the following estimate, for some positive constant independent of the parameter h

$$\|u_h\|_\infty \leq C\|f_h\|_\infty. \tag{128}$$

this implies

* the matrix $\mathcal{L}_h$ is injective, since $f_h = 0$ in [117] implies $u_h = 0$,
* since $\mathcal{L}_h$ is a square matrix, the previous injectivity of $\mathcal{L}_h$ implies the surjectivity,
* estimate [117] yields [117].

Using the computations [53]–[57] combined with the triangle inequality and the fact that $u_0 = 0$, we get

$$|u_i| \leq e^\alpha \|f_h\|_\infty, \ \forall i \in \{1, \ldots, N\}. \tag{129}$$

Which implies

$$\|u_h\|_\infty \leq e^\alpha \|f_h\|_\infty, \tag{130}$$

which means that [128] holds for all $e^\alpha \leq C$.

- Third example In this item, we apply the concepts stated in the second method of Subsection 6.1, on the example of Section 4, that is the finite difference approximation [67]–[68] of [58]–[59]:

– property [103]: let us first set [67]–[68] in the form of [101]. Indeed, $\mathcal{L}_h$ is a square matrix with N lines, and

$$\mathcal{L}_h u_h = \left(-\frac{u_2 - 2u_1 + u_0}{h^2}, \ldots, -\frac{u_N - 2u_{N-1} + u_{N-2}}{h^2}\right)^t, \tag{131}$$

where $u_0 = u_N = 0$, $u_h = (u_1, u_2, \ldots, u_{N-1})^t$, and f_h is the vector of $N - 1$ components

$$f_h = \left(\pi^2 \sin(\pi x_1), \ldots, \pi^2 \sin(\pi x_{N-1})\right)^t. \tag{132}$$

By acting the matrix $\mathcal{L}_h$ on the function u with replacing u_i by $u(x_i)$, for all $i \in \{0, \ldots, N\}$, we get

$$\mathcal{L}_h u = \left(-\frac{u(x_2) - 2u(x_1) + u(x_0)}{h^2}, \ldots, -\frac{u(x_N) - 2u(x_{N-1}) + u(x_{N-2})}{h^2}\right)^t. \tag{133}$$

Using the Taylor expansion [66], we get

$$\begin{aligned}\mathcal{L}_h u &= \left(-\frac{u(x_2)-2u(x_1)+u(x_0)}{h^2},\ldots,-\frac{u(x_N)-2u(x_{N-1})+u(x_{N-2})}{h^2}\right)^t\\ &= \left(\pi^2\sin(\pi x_1)+\beta_1,\ldots,\pi^2\sin(\pi x_{N-1})+\beta_{N-1}\right)^t\\ &= f_h+(\beta_1,\ldots,\beta_{N-1})^t.\end{aligned} \quad [134]$$

Now the function ε_h given by [102] is defined by

$$\varepsilon_h=(\beta_1,\ldots,\beta_N)^t, \quad [135]$$

where β_i, for all $i\in\{1,\ldots,N\}$, are given by [65].

With the assumption [89], that is the fourth derivative of u is bounded uniformly by some positive constant M, we get since $\frac{1}{12}<1$

$$\|\varepsilon_h\|_\infty\le Mh^2, \quad [136]$$

In particular, estimate [136] implies the convergence

$$\|\varepsilon_h\|_\infty\to 0,\ \text{ as } h\to 0. \quad [137]$$

– property [104]: it suffices to prove that if $\mathcal{L}_h u_h=f_h$, where $u_h=(u_1,\ldots,u_{N-1})^t$ and $f_h=(0,\ldots,0)^t$, we have the following estimate, for some positive constant independent of the parameter h

$$\|u_h\|_\infty\le C\|f_h\|_\infty. \quad [138]$$

this implies

* the matrix $\mathcal{L}_h$ is injective, since $f_h=0$ in [117] implies $u_h=0$,
* since $\mathcal{L}_h$ is a square matrix, the previous injectivity of $\mathcal{L}_h$ implies the surjectivity,
* estimate [138] yields [117].

Using the computations [80]–[85] combined with the triangle inequality and the fact that $u_0=u_N=0$ and $k<N$, we get

$$|u_k|\le h^3N^3\|f_h\|_\infty+h^2N^2\|f_h\|_\infty,\ \forall k\in\{2,\ldots,N-1\}, \quad [139]$$

and

$$|u_2|\le h^3N^2\|f_h\|_\infty, \quad [140]$$

which gives, since $Nh=1$ and with the assumption $h\le 1$

$$|u_k|\le 2\|f_h\|_\infty,\ \forall k\in\{2,\ldots,N-1\}, \quad [141]$$

Which implies

$$\|u_h\|_\infty\le 2\|f_h\|_\infty, \quad [142]$$

which means that [138] holds for all $2\le C$.

6.3 A general framework to prove the convergence of the finite difference solution

As we have seen, in general, we do not know both the expression of the exact solution and the finite difference solution. This means that the first method stated in Subsection 6.1 can not be applied in the general case. Whereas the second method of Subsection 6.1 seems to be efficient. This Subsection is devoted to give a "framework" which states the concepts of Subsection 6.1 in some efficient "rule" could be applied whenever we would like to prove the convergence of a given finite difference solution. We will restate here the results of the second method of the Subsection 6.1 but in a more precise manner. As, we have seen that convergence [108] of the finite difference solution u_h towards the exact solution u results from two facts: the first fact is the so called *Consistency* which is the subject of [103], and the second fact is the so called *Stability* which is the subject of [104]. Therefore, the convergence of a given finite difference solution [101] results from the *Consistency* [103] and the Stability [104]. We summarize then this result in the following Theorem:

Theorem 6.1 Let h be a positive parameter, and $\mathcal{L}_h$ be a linear operator from a normed vectorial space $(\mathcal{U}_h; \|\cdot\|_{\mathcal{U}_h})$ into a normed vectorial space $(\mathcal{F}_h; \|\cdot\|_{\mathcal{F}_h})$. Assume that the following properties hold:

- *Stability*: $\mathcal{L}_h$ is invertible and its inverse is bounded by some constant M independent of h:

$$\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h,\mathcal{U}_h)} \le M, \qquad [143]$$

  where

$$\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h,\mathcal{U}_h)} = \sup_{v_h\in\mathcal{F}_h, v_h\neq 0} \frac{\|\mathcal{L}_h^{-1}(v_h)\|_{\mathcal{U}_h}}{\|v_h\|_{\mathcal{F}_h}}.$$

- *Consistency*: Let u_h and $\bar{u}_h$ be two elements from $\mathcal{U}_h$ such that

$$\|\mathcal{L}_h(\bar{u}_h - u_h)\|_{\mathcal{F}_h} \to 0 \text{ as } h \to 0. \qquad [144]$$

Then the following convergence holds:

$$\|\bar{u}_h - u_h\|_{\mathcal{U}_h} \to 0 \text{ as } h \to 0. \qquad [145]$$

Remark 4 The *Stability* given in Theorem 6.1 is equivalent to say, for some constant M independent of h, and for all $r_h \in \mathcal{F}_h$, there exits a unique $q_h \in \mathcal{U}_h$ such that

$$\mathcal{L}_h q_h = r_h. \qquad [146]$$

and

$$\|q_h\|_{\mathcal{U}_h} \le M\|r_h\|_{\mathcal{F}_h}. \qquad [147]$$

Proof The convergence [145] results as follows, thanks to [143]

$$\begin{aligned} \|\bar{u}_h - u_h\|_{\mathcal{U}_h} &= \|\mathcal{L}_h^{-1}\left(\mathcal{L}_h(\bar{u}_h - u_h)\right)\|_{\mathcal{U}_h} \\ &\leq M\|\mathcal{L}_h(\bar{u}_h - u_h)\|_{\mathcal{F}_h}. \end{aligned} \quad [148]$$

Tending h to 0 in the previous inequality and using [144], we get [145] □

6.4 The concept of the convergence order

In the previous Subsection, we provided some sufficient conditions for the convergence of the finite difference solution. This convergence is given in the sense of [145]. It is interesting to measure how it is fast the convergence of the finite difference solution towards the exact solution. More precise, let us consider the following problem:

$$\mathcal{L}u = f, \quad [149]$$

and its finite difference approximation

$$\mathcal{L}u_h = f_h, \quad [150]$$

where h is the parameter mesh discretization.
Let us assume that, there exist two positive constants α and C independent of the parameter mesh discretization h such that

$$\|u - u_h\| \leq Ch^\alpha, \quad [151]$$

where $\|\cdot\|$ is a convenient norm (Some choices of the norm are given in the previous sections, and some discussion of the reasonable choice of these norms will be given below.). As we can see that the estimate [151] yields the convergence of u_h towards u as h tends to 0.
One remarks that for $h \leq 1$, $h^{\alpha_2} \leq h^{\alpha_1}$ for $0 < \alpha_1 < \alpha_2$, one could deduce that as α increases, as the convergence of u_h towards u becomes faster. It is useful then to get α higher.

6.5 How to determine a convergence order of a given finite difference solution?

Theorem 6.1 provides us with some sufficient conditions for the converegence of the finite difference solution towards the exact solution. The following Theorem provides us with some sufficient conditions for a convergence order of the finite difference solution.

THEOREM 6.2 Let h be a positive parameter, and $\mathcal{L}_h$ be a linear operator from a normed vectorial space $(\mathcal{U}_h; \|\cdot\|_{\mathcal{U}_h})$ into a normed vectorial space $(\mathcal{F}_h; \|\cdot\|_{\mathcal{F}_h})$. Assume that the following properties hold:

- *Stability*: $\mathcal{L}_h$ is invertible and its inverse is bounded by some constant M independent of h:

$$\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h, \mathcal{U}_h)} \leq M, \quad [152]$$

where

$$\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h,\mathcal{U}_h)} = \sup_{v_h\in\mathcal{F}_h, v_h\neq 0} \frac{\|\mathcal{L}_h^{-1}(v_h)\|_{\mathcal{U}_h}}{\|v_h\|_{\mathcal{F}_h}}. \qquad [153]$$

- *Consistency*: Let u_h and $\bar{u}_h$ be two elements from $\mathcal{U}_h$ such that, for some two positives constants C and α independent of h

$$\|\mathcal{L}_h(\bar{u}_h - u_h)\|_{\mathcal{F}_h} \le Ch^\alpha. \qquad [154]$$

Then the following convergence holds:

$$\|\bar{u}_h - u_h\|_{\mathcal{U}_h} \le CMh^\alpha. \qquad [155]$$

The Proof of this Theorem follows that one of 6.1.

The following Theorem gives the Theorem [6.2] in the non–linear case of $\mathcal{L}_h$

Theorem 6.3 (Non–linear case) Let h be a positive parameter, and $\mathcal{L}_h$ be an operator from a normed vectorial space $(\mathcal{U}_h; \|\cdot\|_{\mathcal{U}_h})$ into a normed vectorial space $(\mathcal{F}_h; \|\cdot\|_{\mathcal{F}_h})$. Assume that the following properties hold:

- *Stability*: $\mathcal{L}_h$ is invertible and if $\mathcal{L}_h u_h = f_h$ and $\mathcal{L}_h v_h = g_h$ then the following estimate holds, for some constant M independent of h:

$$\|u_h - v_h\|_{\mathcal{U}_h} \le M\|f_h - g_h\|_{\mathcal{F}_h}. \qquad [156]$$

- *Consistency*: Let u_h and $\bar{u}_h$ be two elements from $\mathcal{U}_h$ such that, for some two positives constants C and α independent of h

$$\|\mathcal{L}_h(\bar{u}_h - u_h)\|_{\mathcal{F}_h} \le Ch^\alpha. \qquad [157]$$

Then the following convergence holds:

$$\|\bar{u}_h - u_h\|_{\mathcal{U}_h} \le CMh^\alpha. \qquad [158]$$

Remark 5 (Theorem 6.3 generalizes 6.3) Equality [156] generalizes [152] when we put $v_h = 0_{\mathcal{U}_h}$ and $g_h = 0_{\mathcal{F}_h}$ where $0_{\mathcal{U}_h}$ and $0_{\mathcal{F}_h}$; then [156] gives $\|u_h\|_{\mathcal{U}_h} \le M\|f_h\|_{\mathcal{F}_h}$ which means that $\|\mathcal{L}_h^{-1}(f_h)\|_{\mathcal{U}_h} \le M\|f_h\|_{\mathcal{F}_h}$. This yields that $\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h,\mathcal{U}_h)} \le M$, according to the definition [153] of the norm $\|\mathcal{L}_h^{-1}\|_{\mathcal{L}(\mathcal{F}_h,\mathcal{U}_h)}$. Therefore, Theorem 6.3 generalizes Theorem 6.2

6.6 Some examples of the finite difference approximation

In this Subsection, we quote some examples of the finite difference approximation of ordinary differential equations as well as of partial differential equations. We will apply, in these examples, Theorem 6.2 in order to determine a convergence order of these finite difference approximations.

- First example Let us consider the following problem

$$-u''(x) + (1+x^2)u(x) = \sqrt{1+x},\ x \in (0,1), \tag{159}$$

with the Dirichlet boundary conditions

$$u(0) = u(1) = 0. \tag{160}$$

Let h be a positive parameter which is expected to approah 0. We introduce the finite difference discretization $x_i = ih$, for all $i \in \{0, \ldots, N\}$ where $x_0 = 0$ and $x_N = 1$.
The finite difference approximation we suggest to approximate [159]–[160]:

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + (1+x_i^2)u_i = \sqrt{1+x_i},\ i \in \{1, \ldots, N-1\}, \tag{161}$$

with

$$u_0 = u_N = 0. \tag{162}$$

In order to prove the existence, uniqueness, and convergence of the finite difference solution $(u_i)_0^N$ of [161]–[162] towards the exact solution u of, [159]–[160], we will apply Theorem 6.2.

  - *Stablity* : We could set [161]–[162] in the following form:

$$\mathcal{L}_h u_h = f_h, \tag{163}$$

where

$$u_h = (u_1, \ldots, u_{N-1})^t, \tag{164}$$

and $\mathcal{L}_h$ is the square matrix of $N-1$ lines defined by

$$\mathcal{L}_h u_h = \left(-\frac{u_2 - 2u_1 + u_0}{h^2} + (1+x_1^2)u_1, \ldots, -\frac{u_N - 2u_{N-1} + u_{N-2}}{h^2} + (1+x_{N-1}^2)u_{N-1}\right)^t, \tag{165}$$

with $u_0 = u_N = 0$, and the second member f_h is defined by

$$f_h = \left(\sqrt{1+x_1}, \ldots, \sqrt{1+x_{N-1}}\right)^t. \tag{166}$$

To prove [152], we first prove that there exists a constant a positive constant M independent of h such that for any given vector $f_h = (f_1, \ldots, f_{N-1})^t$ and for any possible solution $u_h = (u_1, \ldots, f_{N-1})^t$ of $\mathcal{L}u_h = f_h$, the following estimate

$$\max(|u_1|, \ldots, |u_{N-1}|) \le M \max(|f_1|, \ldots, |f_{N-1}|) \tag{167}$$

Indeed, estimate [167] yields:

    * injectivity of $\mathcal{L}_h$ in the sense: $\mathcal{L}_h u_h = 0$ implies, by replacing $f_h = 0$ in [167], $u_h = 0$
    * since $\mathcal{L}_h$ is a square matrix, then this last injectivity implies the surjectivity in the sense that for all $f_h = (f_1, \ldots, f_{N-1})^t$ there exists a unique (this uniqueness is the subject of the previous item) $u_h = (u_1, \ldots, f_{N-1})^t$ such that $\mathcal{L}u_h = f_h$,

* estimate [167] gives estimate [152]

Let us first write $\mathcal{L}u_h = f_h$ in the following form, thanks to [165]

$$-\frac{1}{h^2}u_{i+1} + \frac{2+h^2(1+x_i^2)}{h^2}u_i - \frac{1}{h^2}u_{i-1} = f_i,\ i \in \{1,\dots,N-1\}, \qquad [168]$$

with $u_0 = u_N = 0$.

Assume that there exists $k \in \{1,\dots,N-1\}$ such that $|u_k| = \max(|u_1|,\dots,|u_{N-1}|)$, and writing [168] when $i = k$

$$-\frac{1}{h^2}u_{k+1} + \frac{2+h^2(1+x_k^2)}{h^2}u_k - \frac{1}{h^2}u_{k-1} = f_k, \qquad [169]$$

which implies that

$$\frac{2+h^2(1+x_k^2)}{h^2}u_k = f_k + \frac{1}{h^2}u_{k-1} + \frac{1}{h^2}u_{k+1}, \qquad [170]$$

this with the triangle inequality and $|u_{k-1}|, |u_{k-1}| \le |u_k|$ implies

$$\frac{2+h^2(1+x_k^2)}{h^2}|u_k| \le |f_k| + \frac{1}{h^2}|u_k| + \frac{1}{h^2}|u_k|. \qquad [171]$$

Which implies in turn that

$$(1+x_k^2)|u_k| \le |f_k|. \qquad [172]$$

This yields, since $1 + x_k^2 > 1$

$$|u_k| \le |f_k|. \qquad [173]$$

Since $|u_k| = \max(|u_1|,\dots,|u_{N-1}|)$ and $|f_k| \le \max(|f_1|,\dots,|f_{N-1}|)$, estimate [173] implies

$$\max(|u_1|,\dots,|u_{N-1}|) \le \max(|f_1|,\dots,|f_{N-1}|). \qquad [174]$$

– Consistency By acting the matrix $\mathcal{L}_h$ on the vector $[u]_h = (u(x_1),\dots,u(x_{N-1}))^t$ with $u(x_0) = u(x_N) = 0$, we get

$$\begin{aligned}\mathcal{L}_h[u]_h &= (-\frac{u(x_2)-2u(x_1)+u(x_0)}{h^2} + (1+x_1^2)u(x_1),\dots,-\frac{u(x_N)-2u(x_{N-1})+u(x_{N-2})}{h^2}\\ &+ (1+x_{N-1}^2)u(xN-1))^t. \qquad [175]\end{aligned}$$

Using the Taylor expansion [66] and equation [159], we get

$$\begin{aligned}\mathcal{L}_h[u]_h &= \left(\sqrt{1+x_1}+\beta_1,\dots,\sqrt{1+x_{N-1}}+\beta_{N-1}\right)^t\\ &= f_h + (\beta_1,\dots,\beta_N)^t. \qquad [176]\end{aligned}$$

Substracting [163] from [176], we get

$$\mathcal{L}_h([u]_h - u_h) = (\beta_1,\dots,\beta_{N-1})^t. \qquad [177]$$

where β_i, for all $i \in \{1,\dots,N\}$, are given by [65].

With the assumption [89], that is the fourth derivative of u is bounded uniformly by some positive constant M, we get since $\frac{1}{12} < 1$

$$\|\mathcal{L}_h([u]_h - u_h)\| \le Mh^2, \qquad [178]$$

where the norm $\|\cdot\|$ is the norm defined by

$$\| (s_1, \ldots, s_{N-1})^t \| = \max (|s_1|, \ldots, |s_{N-1}|) . \quad [179]$$

Using now Theorem [6.2], we get

$$\|[u]_h - u_h\| \le Mh^2. \quad [180]$$

Remark 6 (Other Stability for [163]–[166]) For the discrete problem [163]–[166], we have proven the Stability [167]. It is also possible to prove a Stability in other norm, it is given in Remark 3. This norm can be viewed as in H_0^1 norm, that the norm defined by

$$\|u_h\|^2_{H_0^1} = \sum_{i=0}^{N-1} h \left(\frac{u_{i+1} - u_i}{h} \right)^2 . \quad [181]$$

Indeed, $\mathcal{L}_h u_h = f_h$, with $f_h = (f_1, \ldots, f_{N-1})^t$ means that

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + (1 + x_i^2)u_i = f_i,\ i \in \{1, \ldots, N-1\}, \quad [182]$$

which could be written as

$$-\frac{u_{i+1} - u_i}{h} + \frac{u_i - u_{i+1}}{h} + (1 + x_i^2)u_i = hf_i,\ i \in \{1, \ldots, N-1\}. \quad [183]$$

Multiplying both sides of [183] by u_i, summing over $i \in \{1, \ldots, N-1\}$, reording the sum, and using the fact that $u_0 = u_N = 0$, we get

$$\|u_h\|^2_{H_0^1} + \sum_{i=1}^{N-1} (1 + x_i^2)u_i^2 = h \sum_{i=1}^{N-1} f_i u_i. \quad [184]$$

Since $(1 + x_i^2)u_i^2 \ge 0$, [184] implies

$$\|u_h\|^2_{H_0^1} \le h \sum_{i=1}^{N-1} f_i u_i. \quad [185]$$

The right hand side of the previous inequality could be estimated as

$$h \sum_{i=1}^{N-1} f_i u_i \le \max (|f_1|, \ldots, |f_{N-1}|) \max (|u_1|, \ldots, |u_{N-1}|) . \quad [186]$$

Let us assume that, for some $k \in \{1, \ldots, N-1\}$

$$\max (|u_1|, \ldots, |u_{N-1}|) = |u_k| \quad [187]$$

We have, thanks to the Cauchy Schwarz inequality since $u_0 = 0$

$$\begin{aligned} |u_k| &= |\sum_1^k (u_j - u_{j-1})| \\ &\le \left(\sum_1^k \frac{(u_j - u_{j-1})^2}{h} \right)^{\frac{1}{2}} \left(\sum_1^k h \right)^{\frac{1}{2}} \\ &\le \|u_h\|_{H_0^1} (Nh)^{\frac{1}{2}} \\ &= \|u_h\|_{H_0^1}. \end{aligned} \quad [188]$$

This with [185]–[187] imply

$$\|u_h\|_{H_0^1} \leq \max\left(|f_1|, \ldots, |f_{N-1}|\right). \qquad [189]$$

- Second example, see [EYM 00, Section 8, Pages 749–754] Let us consider the following semi–linear equation:

$$-u_{xx}(x) = f(x, u(x)),\ x \in (0,1) \qquad [190]$$

with the boundary condition

$$u(0) = 0. \qquad [191]$$

For the sake of simplicity, we assume that the function $f(x,s)$ is continuous with respect to both variables x and s. We assume in addition that

$$f \in L^\infty\left((0,1) \times \mathbb{R}\right) \qquad [192]$$

A weak formulation for equation [190]–[191] may be given by: find $u \in H_0^1(0,1)$ such that

$$\int_0^1 u_x(x)\varphi_x(x)dx = \int_0^1 f(x, u(x))\varphi_x(x)dx,\ \forall \varphi \in H_0^1(0,1), \qquad [193]$$

where $H_0^1(0,1)$ denotes, as usual, the space $v \in L^2(0,1)$ such that $v_x \in L^2(0,1)$ and $v(1) = v(0) = 0$. The existence of at least one solution for [193] could be proven thanks, e.g., to Schauder's fixed point theorem or by using the convergence of the numerical schemes.
Inspiring the ideas of Section 4, we suggest the following finite difference scheme:

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f(x_i, u_i),\ \forall i \in \{1, \ldots, N-1\}, \qquad [194]$$

where u_i is an approximation of $u(x_i)$, for all $i \in \{0, \ldots, N\}$. Since $u(0) = u(1) = 0$, we chose

$$u_0 = u_N = 0. \qquad [195]$$

First step:

We first jutify the existence of a vector $(u_i)_{i=1}^N$ satisfying [194] with $u_0 = u_N = 0$.
For this purpose, we apply the so-called Brouwer's theorem. Let

$$M = \|f\|_{L^\infty((0,1)\times\mathbb{R})}. \qquad [196]$$

Let $V = (v_1, \ldots, v_{N-1}) \in \mathbb{R}^{N-1}$, there exists a unique solution $U = (u_1, \ldots, u_{N-1}) \in \mathbb{R}^{N-1}$ of [194]–[195] by replacing $f(x_i, u_i)$ with $f(x_i, v_i)$ in the right hand side of [194]. One sets, $\mathcal{F}(U) = V$.
So $\mathcal{F}$ is continuous since $\mathbb{R}^{N-1}$ is a finite dimensional space.
Multiplying both sides of [194] by u_i, summing over $i \in \{1, \ldots, N-1\}$ wet get

$$\frac{1}{h^2}\left(-\sum_{i=1}^{N-1}\left(u_{i+1} - u_i\right)u_i + \sum_{i=1}^{N-1}\left(u_i - u_{i-1}\right)u_i\right) = \sum_{i=1}^{N-1} f(x_i, v_i)u_i. \qquad [197]$$

Re-ordering the sum of the second term in left hand side of the previous equality and using the discrete "boundary" condition [195], we get

$$\begin{aligned}\sum_{i=1}^{N-1}(u_i-u_{i-1})\,u_i &= \sum_{i=1}^{N}(u_i-u_{i-1})\,u_i\\ &= \sum_{i=0}^{N-1}(u_{i+1}-u_i)\,u_{i+1}.\end{aligned} \tag{198}$$

On the other hand, the first term in the left hand side of [197] could be written as, since $u_0=0$

$$\sum_{i=1}^{N-1}(u_{i+1}-u_i)\,u_i=\sum_{i=0}^{N-1}(u_{i+1}-u_i)\,u_i. \tag{199}$$

Combining now [197]–[199], we get

$$\frac{1}{h^2}\left(\sum_{i=0}^{N-1}(u_{i+1}-u_i)^2\right)=\sum_{i=1}^{N-1}f(x_i,v_i)u_i, \tag{200}$$

which is equivalent to

$$\sum_{i=0}^{N-1}\frac{(u_{i+1}-u_i)^2}{h}=\sum_{i=1}^{N-1}f(x_i,v_i)hu_i. \tag{201}$$

Using now [196], the previous equality yields

$$\sum_{i=0}^{N-1}\frac{(u_{i+1}-u_i)^2}{h}\le M\sum_{i=1}^{N-1}h|u_i|. \tag{202}$$

Which implies, using the fact that $\sum_{i=1}^{N-1}h<1$

$$\sum_{i=0}^{N-1}\frac{(u_{i+1}-u_i)^2}{h}\le M\max\left(|u_1|,\ldots,|u_{N-1}|\right). \tag{203}$$

Using inequality [188] and definition [181], inequality [203] implies that

$$\|u_h\|_{H_0^1}\le M, \tag{204}$$

where $u_h=(u_1,\ldots,u_{N-1})$.

Now the application $\mathcal{F}$ defined above is continuous, and taking in $\mathbb{R}^{N-1}$ the norm $\|V\|_{H_0^1}$ defined by [181], with $V=(v_1,\ldots,v_{N-1})$ and $v_0=v_N=0$.

Estimate [204] yields $\mathcal{F}(B_M)\subset B_M$. Thanks to Brouwer fixed point theorem, $\mathcal{F}$ has a fixed point, and this fixed point is a solution for [194]–[195].

Second step: We assume, for instance, in order to get a convergence order for the finite difference solution [194]–[195], that $f\in\mathcal{C}^1([0,1]\times\mathbb{R},\mathbb{R})$ and the following condition on the function f holds, for some $\gamma\in(0,1)$ such that

$$\left(f(x,s)-f(x,t)\right)(s-t)\le\gamma(s-t)^2,\ \forall(x,s)\in[0,1]\times\mathbb{R}. \tag{205}$$

Using [64] and equation [190], we get

$$-\frac{u(x_{i+1})-2u(x_i)+u(x_{i-1})}{h^2}=f(x_i,u(x_i))-\beta_i,\ \forall i\in\{1,\ldots,N-1\}, \tag{206}$$

where β_i is given by [65].

Substracting [194] from [206], we get

$$-\frac{e_{i+1}-2e_i+e_{i-1}}{h^2}=f(x_i,u(x_i))-f(x_i,u_i)+\beta_i,\ \forall i\in\{1,\ldots,N-1\},\qquad [207]$$

where $e_i=u(x_i)-u_i$, for all $i\in\{0,\ldots,N\}$.

Multiplying both sides of [207] by e_i and using techniques used in [197]–[201], we get

$$\|e_h\|^2_{H^1_0}=\sum_{i=1}^{N-1}h\left(f(x_i,u(x_i))-f(x_i,u_i)\right)e_i+\frac{h^2}{12}\bar{M}\|e_h\|_{H^1_0},\qquad [208]$$

where $e_h=(e_1,\ldots,e_{N-1})$ and $\bar{M}=\max\limits_{x\in[0,1]}|u_{xx}(x)|$.

Using now [205], we get

$$\left(f(x_i,u(x_i))-f(x_i,u_i)\right)e_i\le\gamma e_i^2,\ \forall i\in\{1,\ldots,N-1\},\qquad [209]$$

and therefore, thanks to [188] and $\sum_{i=1}^{N-1}h<1$

$$\sum_{i=1}^{N-1}h\left(f(x_i,u(x_i))-f(x_i,u_i)\right)e_i\le\gamma\|e_h\|^2_{H^1_0},\ \forall i\in\{1,\ldots,N-1\},\qquad [210]$$

Equation [208] becomes then, thanks to [207]–[210]

$$(1-\gamma)\|e_h\|^2_{H^1_0}\le\frac{h^2}{12}\bar{M}\|e_h,\|_{H^1_0}.\qquad [211]$$

which implies that

$$\|e_h\|_{H^1_0}\le\frac{\bar{M}}{1-\gamma}h^2.\qquad [212]$$

Estimate [212] also yields, thanks to [188]

$$\max\left(|e_1|,\ldots,|e_{N-1}|\right)\le\frac{\bar{M}}{1-\gamma}h^2.\qquad [213]$$

7 Some simulations in Scilab

This section is devoted to justify numerically the theoretical results given in sections 2, 3 and 4. The following tables show:

- Error: the error is defined by $\max\left(|u(x_1)-u_1|,\ldots,|u(x_N)-u_N|\right)$, in cases of the examples of section 2 and 3, and the error is defined by $\max\left(|u(x_1)-u_1|,\ldots,|u(x_{N-1})-u_{N-1}|\right)$ in case of the example of 4.
- Convergence order: the convergence order is computed, as usual, thanks to the following rule:

$$\frac{\log(E(n))-\log(E(n+1))}{\log(2)},\qquad [214]$$

  where $E(n)$ is the error, defined in the previous item, corresponding to $h=\frac{1}{2^n}$.
- Simulations: We will use the following simulations in Scilab:

– Examples of Section 2 (since the simulations of example 3 is similar to those of example given 2):

  * N % *The number of mesh points*;
  * $M = N - 1$ % *The dimension of the unknown vector*;
  * $h = 1/N$ % *The mesh size*;
  * $X = [h : h : 1]$ % *The mesh points in which we approximate u*
  * $UX = X.^2$ % *The exact solution*
  * for $i = 1 : M$, $U = i * (i-1) * h^2$ % *The finite difference solution*
  * for $i = 1 : M$, $E(i) = abs(U(i) - UX(i))$ % *The absolute values of the components of the vector error*
  * $error = \max(E)$ % *The maximum value of the components of the vector error*

– Examples of Section 4

  * N % *The number of mesh points*;
  * $M = N - 1$ % *The dimension of the matrix*;
  * $h = 1/N$ % *The mesh size*;
  * $X = [h : h : 1 - h]$ % *The mesh points in which we approximate u*
  * $UX = \sin(\%\pi * X)$ % *The exact solution*
  * for $i = 1 : M$, $A(i, j) = 0; end; end;$ % *Initialization of the matrix A*
  * for $i = 1 : M$, $A(i, i) = 2; end; end;$ % *Initialization of the matrix A*
  * for $i = 1 : M - 1$, $A(i, i+1) = -1; end; end;$
  * for $i = 1 : M - 1$, $A(i+1, i) = -1; end; end;$
  * $U = h^2 * \%\pi^2 * inv(A) * (\sin(\%\pi * X))'$
  * for $i = 1 : M$, $E(i) = abs(U(i) - UX(i))$ % *The absolute values of the components of the vector error*
  * $error = \max(E)$ % *The maximum value of the components of the vector error*

To justify rule [214], let us assume that the order of the finite difference approximation is α and then we could write, for some case, where $h = \frac{1}{2^n}$

$$E(n) = \{\frac{1}{2^n}\}^\alpha, \qquad [215]$$

which is equivalent to

$$E(n) = \frac{1}{2^{\alpha n}}. \qquad [216]$$

Therefore

$$\log(E(n)) = -\alpha n \log(2). \qquad [217]$$

Substituting n by $n+1$ in [217], we get

$$\log(E(n+1)) = -\alpha(n+1)\log(2). \qquad [218]$$

Substracting [217] from [218] and dividing the result by log(2), we get

$$\alpha = \frac{\log(E(n)) - \log(E(n+1))}{\log(2)}. \qquad [219]$$

We will remark that as h decreases to approach zero, as the error decreases to approach zero.

7.1 Simulations for the example given in section 2

h	Error	Order
1/32	0.3125	-
1/64	0.015625	1.
1/128	0.0078125	1.
1/256	0.0039062	1.
1/512	0.0019531	0.9999261
1/1024	0.0009766	0.9999261
1/2048	0.0004883	0.9998818

The numerical results given in the previous table justify well theoretical results given in Section 2. Indeed, thanks to [22], the error is bounded by

$$\max_{i\in\{0,N\}} |u(x_i) - u_i| \leq h. \qquad [220]$$

We have, since $x_N = 1$ and $x_{N-1} = 1 - h$

$$\begin{aligned} |u(x_N) - u_N| &= |u(1) - u_N| \\ &= |1 - x_N x_{N-1}| \\ &= |1 - (1-h)| \\ &= h. \qquad [221] \end{aligned}$$

This with [220] implies that

$$\max_{i\in\{0,N\}} |u(x_i) - u_i| = h. \qquad [222]$$

This last result is justified by the previous table by comparing the values of h in the first column and the values of the error in second column.

7.2 Simulations for the example given in section 3

h	Error	Order
1/8	0.1524973	-
1/16	0.0803533	0.9243545
1/32	0.0412917	0.9605055
1/64	0.0209369	0.9798040
1/128	0.0105428	0.9897898
1/256	0.0052902	0.9948639
1/512	0.0026498	0.9974388
1/1024	0.0013261	0.9986939

The previous table shows that the finite difference solution [36]–[37] converges towards the exact solution of [32]–[33] by order h.

7.3 Simulations for the example given in section 4

h	Error	Order
1/8	0.0129507	-
1/16	0.0032190	2.0083456
1/32	0.0008036	2.0020631
1/64	0.0002008	2.0007183
1/128	0.0000502	2.
1/256	0.0000125	2.0057593
1/512	0.0000031	2.011588
1/1024	0.0000008	1.9541963

8 Finite difference methods for higher dimension equations

So far we considered the finite difference approximation for one dimension equation. We will devote this subsection to stationary two dimensional equations.

8.1 A first example

Let us consider the following two dimensional equation:

$$-\Delta u(x,y) = \varphi(x,y),\ (x,y) \in \Omega = (0,1)^2, \tag{223}$$

with Dirichlet boundary condition

$$u(x,y) = \psi(x,y),\ (x,y) \in \partial\Omega, \tag{224}$$

where Δ denotes the Laplace operator:

$$\Delta\, u(x,y) = \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) \tag{225}$$

The finite difference approximation for problem [223]–[224] can be performed via the following steps:

1. finite difference mesh For a given positive *parameter* $h = \frac{1}{N}$, with $N \in \mathbb{N}$, is expected to tend towards zero, we consider the following set of mesh points:

$$\mathcal{D}_h = \{(mh, nh),\ (m,n) \in \{0,\ldots,N\} \times \{0,\ldots,N\}\}. \tag{226}$$

   we denote by

$$(x_m, y_n) = (mh, nh),\ \forall\ (m,n) \in \{0,\ldots,N\} \times \{0,\ldots,N\}$$

2. finite difference scheme: we consider the following scheme: find $\{u_{m,n};\ (m,n) \in \{1,\ldots,N-1\} \times \{1,\ldots,N-1\}\}$ such that, for all $(m,n) \in \{1,\ldots,N-1\} \times \{1,\ldots,N-1\}$

$$-\frac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{h^2} - \frac{u_{m,n+1} - 2u_{m,n} + u_{m,n-1}}{h^2} = \varphi(x_m, y_n), \tag{227}$$

   where, according with the boundary condition [224], we set

$$u_{m,0} = \psi(mh, 0),\ \forall\ m \in \{0,\ldots,N\}, \tag{228}$$

$$u_{m,N} = \psi(mh, 1),\ \forall\ m \in \{0,\ldots,N\}, \tag{229}$$

$$u_{0,n} = \psi(0, nh),\ \forall\ n \in \{0,\ldots,N\}, \tag{230}$$

$$u_{N,n} = \psi(1, nh),\ \forall\ n \in \{0,\ldots,N\}. \tag{231}$$

The analysis of the finite difference scheme [227]–[231] could be performed via the following steps:

1. first step: existence of $\{u_{m,n};\ (m,n) \in \{0,\ldots,N\} \times \{0,\ldots,N\}\}$ satisfying [227]–[231],

2. first step: convergence $\{u_{m,n};\ (m,n) \in \{0,\ldots,N\} \times \{0,\ldots,N\}\}$ towards the exact solutio u of [223]–[224] in some sense.

Let us denote

$$u_h = \left(u_{m,n}\right)_{(m,n)\in\{0,\ldots,N\}\times\{0,\ldots,N\}\}}. \tag{232}$$

1. Existence and uniqueness of the solution u_h defined by [227]–[231]: we will such existence and uniqueness by using two methods:

   (a) first method: Let us assume that there are two solutions $u_h^1 = \left(u_{m,n}^1\right)_{(m,n)\in\{0,\ldots,N\}\times\{0,\ldots,N\}}$ and $u_h^2 = \left(u_{m,n}^2\right)_{(m,n)\in\{0,\ldots,N\}\times\{0,\ldots,N\}}$ for [227]–[231] and consider $\bar{u}_h = u_h^1 - u_h^2$; the vector $\bar{u}_h$ is satisfying, for all $(m,n) \in \{1,\ldots,N-1\} \times \{1,\ldots,N-1\}$

$$-\frac{\bar{u}_{m+1,n} - 2\bar{u}_{m,n} + \bar{u}_{m-1,n}}{h^2} - \frac{\bar{u}_{m,n+1} - 2\bar{u}_{m,n} + \bar{u}_{m,n-1}}{h^2} = 0, \tag{233}$$

with, thanks to the boundary condition [224], for all $(m,n) \in \{1,\dots,N-1\}\times\{0,\dots,N\}$

$$\bar{u}_{m,0} = \bar{u}_{m,N} = \bar{u}_{0,n} = \bar{u}_{N,n} = 0,\ \forall\ m \in \{0,\dots,N\}, \tag{234}$$

Multiplying both sides of [233] by $h^2 u_{m,n}$, summing over $(m,n) \in \{1,\dots,N-1\} \times \{1,\dots,N-1\}$, and re–ordering the sum, we get

$$\sum_{n=1}^{N-1}\sum_{m=0}^{N-1}(\bar{u}_{m+1,n}-\bar{u}_{m,n})^2 + \sum_{m=}^{N-1}\sum_{n=0}^{N-1}(\bar{u}_{m,n+1}-\bar{u}_{m,n})^2 = 0. \tag{235}$$

This implies that, for all $(m,n) \in \{0,\dots,N-1\} \times \{1,\dots,N-1\}$

$$\bar{u}_{m+1,n} = \bar{u}_{m,n}, \tag{236}$$

and, for all $(m,n) \in \{1,\dots,N-1\} \times \{0,\dots,N-1\}$

$$\bar{u}_{m,n+1} = \bar{u}_{m,n}. \tag{237}$$

These two previous equations with [234] imply that, for all $(m,n) \in \{0,\dots,N\} \times \{0,\dots,N\}$

$$\bar{u}_{m,n} = 0. \tag{238}$$

This implies that $u_h^1 = u_h^2$, which means the uniqueness of the solution of [227]–[231]. We use this uniqueness to prove the existence of solution for [227]–[231]. Indeed, [227] is a linear system of $(N-1)^2$ uknowns and $(N-1)^2$. Thus the uniqueness implies the existence.

(b) second method:discrete maximum principle we use here the so called *discrete maximum principle* whose its statement is:

LEMMA 8.1 (Discrete maximum principle) Let $\mathcal{D}_h$ be the discretization given by [226]. Consider the discrete operator $\mathcal{L}_h$ defined by: for a given vector $u_h = (u_{m,n})_{(m,n)\in\{0,\dots,N\}\times\{0,\dots,N\}}$, we define $\mathcal{L}_h\, u_h$ as the discrete function defined on $\mathcal{D}_h$ and takes their values as follows:

$$\mathcal{L}_h\, u_h(x_m,y_n) \begin{cases} -\frac{u_{m+1,n}-2u_{m,n}+u_{m-1,n}}{h^2} - \frac{u_{m,n+1}-2u_{m,n}+u_{m,n-1}}{h^2}, (x_m,y_n) \in \Omega_h \\ u_{m,n},\ (x_m,y_n) \in \mathcal{D}_h \diagdown \Omega_h, \end{cases} \tag{239}$$

where Ω_h denotes the set of the interior mesh points, that is

$$\Omega_h = \{(x_m,y_n) \in \Omega\} = \{1,\dots,N-1\} \times \{1,\dots,N-1\}, \tag{240}$$

and $\mathcal{D}_h \diagdown \Omega_h$ denotes the mesh points which locate on the boundary of Ω.
Assume that, for all $(m,n) \in \{1,\dots,N-1\} \times \{1,\dots,N-1\}$

$$\mathcal{L}_h\, u_h(x_m,y_n) \le\ 0. \tag{241}$$

Then u_h reaches its maximum at least on some point $u(x_i,y_j)$ such that $(x_i,y_j) \in \partial\Omega$.

Proof Assume the contrary. This means that the maximum of u_h could be only reached on the interior mesh points.

Consider the set of the interior mesh points where the maximum of u_h is reached:

$$\gamma_h = \{(x_r, y_s) \in \Omega_h : u_{r,s} = \max\{u_{m,n}, (m,n) \in \Omega_h\}\}. \qquad [242]$$

and consider

$$i = \max\{m; (x_m, y_n) \in \gamma_h\}. \qquad [243]$$

Let then $j \in \{1, \ldots, N-1\}$ such that $(x_i, y_j) \in \gamma_h$.

Writing [241] when $(m,n) = (i,j)$ and multiplying the result by $-h^2$, we get

$$(\, u_{i+1,j} - u_{i,j}) + (\, u_{i-1,j} - u_{i,j}) + (\, u_{i,j+1} - u_{i,j}) + (\, u_{i,j-1} - u_{i,j}) \geq \, 0. \qquad [244]$$

The left hand side of the previous expression contains four negative terms, and the first term is non positive else $u_{i+1,j}$ is also maximum and then $(i+1, j) \in \gamma_h$ because u_h could reach its maximum only on interior points. $u_{i+1,j}$ is maximum is a contradiction with [243]. □

The following lemma is also required for the question of existence and uniqueness of the solution of [227]–[231].

Lemma 8.2 Let $\mathcal{D}_h$ be the discretization given by [226]. Consider the discrete operator $\mathcal{L}_h$ defined by: for a given vector $u_h = (\, u_{m,n})_{(m,n) \in \{0,\ldots,N\} \times \{0,\ldots,N\}}$, we define $\mathcal{L}_h \, u_h$ as the discrete function defined on $\mathcal{D}_h$ and takes their values as it is defined in [239]–[240]. Assume that, for all $(m,n) \in \{1, \ldots, N-1\} \times \{1, \ldots, N-1\}$

$$\mathcal{L}_h \, u_h(x_m, y_n) \geq \, 0. \qquad [245]$$

Then u_h reachs its minimum at least on some point $u(x_i, y_j)$ such that $(x_i, y_j) \in \partial\Omega$.

Assume now that there two solutions u_h^1 and u_h^2 for [227]–[231]. Therefore $\bar{u}_h = u_h^1 - u_h^2$ satisfies

$$(\, \mathcal{L}_h \, \bar{u}_h)_{(m,n)} = \, 0, \qquad [246]$$

and

$$\bar{u}_{0,n} = \bar{u}_{N,n} = \bar{u}_{m,0} = \bar{u}_{m,N} = 0, \; \forall \, (m,n) \in \{0, \ldots, N\} \times \{0, \ldots, N\}. \qquad [247]$$

Thanks to Lemma 8.1, $\bar{u}_h$ can reach its maximum at least on some (i,j) such that $(x_i, y_j) \in \partial\Omega$. One knows that the value of $\bar{u}_h$ on (i,j) is zero, thanks to [247], one could deduces that $\bar{u}_h$ is negative. By the same way, namely using Lemma 8.2 and [247], we deduce that $\bar{u}_h$ is positive. Therefore $\bar{u}_h = 0$ and then $u_h^1 = u_h^2$ which proves the uniqueness of the solution of [227]–[231].

Now to justify the existence of the solution of [227]–[231], we use the uniqueness of the solution of [227]–[231]. Indeed, [227]–[231] could be written as a linear system of $(N-1)^2$ unknowns, namely $\{u_{m,n}; (m,n) \in \{1,\ldots,N-1\} \times \{1,\ldots,N-1\}$ and $(N-1)^2$ equations. Since we have the uniqueness of the solution of [227]–[231] , then we have the existence of a solution for [227]–[231].

2. Stability

For a given vector $\varphi_h = (\varphi_{m,n})_{(m,n)\in\{1,\ldots,N-1\}\times\{1,\ldots,N-1\}}$, we consider the vector $v_h = (v_{m,n})_{(x_m,y_n)\in\mathcal{D}_h}$ as the solution of

and

$$\mathcal{L}_h\, v_h(x_m, y_n) = \varphi_{m,n}, \;, \; \forall (x_m, y_n) \in \Omega_h, \tag{248}$$

with

$$v_{m,n} = 0, \; \forall\, (x_m, y_n) \in \mathcal{D}_h \diagdown \Omega_h, \tag{249}$$

where $\mathcal{L}_h$ is defined by [239].

Let us consider the positive quantity $\|\varphi\|$ defined by:

$$\|\varphi\| = \max_{(m,n)} |\varphi_{mn}|, \tag{250}$$

Let us consider the following function

$$\mathcal{P}(x, y) = \frac{1}{4}\left(3 - (x^2 + y^2)\right)\|\varphi\|. \tag{251}$$

and its approximation $\mathcal{P}_h = (\mathcal{P}_{m,n})_{(x_m,y_n)\in\mathcal{D}_h}$ given by

$$\mathcal{L}_h \mathcal{P}_h(x_m, y_n) = -\Delta \mathcal{P}_{x_m, y_n}, \; \forall (x_m, y_n) \in \Omega_h, \tag{252}$$

with

$$\mathcal{P}_{m,n} = \mathcal{P} x_m, y_n, \; (x_m, y_n) \in \mathcal{D}_h \diagdown \Omega_h. \tag{253}$$

Using a Taylor expansion, we get

$$\mathcal{L}_h\, \mathcal{P}(x_m, y_n) = -\Delta \mathcal{P}(x_m, y_n), \; \forall (x_m, y_n) \in \Omega_h. \tag{254}$$

This with previous items of uniqueness leads to

$$\mathcal{P}(x_m, y_n) = \mathcal{P}_{m,n}, \; \forall (x_m, y_n) \in \mathcal{D}_h. \tag{255}$$

Since

$$-\,\Delta \mathcal{P}(x_m, y_n) = \|\varphi\|, \tag{256}$$

then [252] becomes

$$\mathcal{L}_h \mathcal{P}_h(x_m, y_n) = \|\varphi\|, \; \forall (x_m, y_n) \in \Omega_h. \tag{257}$$

Substracting [257] from [248] yields that

$$\mathcal{L}_h\, (v_h - \mathcal{P}_h)(x_m, y_n) = \varphi_{m,n} - \|\varphi\|, \; \forall (x_m, y_n) \in \Omega_h. \tag{258}$$

Since $\varphi_{m,n} - \| \varphi\| \leq 0$, for all $(x_m, y_n) \in \Omega_h$, then thanks to Lemma 8.1, $v_h - \mathcal{P}_h$ takes its maximum at least on some boundary mesh point $(x_i, y_j) \in \partial\Omega$. Therefore, using [249]

$$\max (v_h - \mathcal{P}_h) \leq -\frac{1}{4} \left(3 - (x_i^2 + y_j^2)\right) \| \varphi\|. \qquad [259]$$

One remarks that

$$x_i^2 + y_j^2 \leq 2, \qquad [260]$$

one could deduce that

$$\frac{1}{4} \left(3 - (x_i^2 + y_j^2)\right) \| \varphi\| \leq \| \varphi\| \geq 0 \qquad [261]$$

which implies that

$$-\frac{1}{4} \left(3 - (x_i^2 + y_j^2)\right) \| \varphi\| \leq 0. \qquad [262]$$

This with [259] yields that

$$\max (v_h - \mathcal{P}_h) \leq 0, \qquad [263]$$

which means that

$$v_{m,n} \leq \mathcal{P}_{m,n}, \ \forall (x_m, y_n) \in \mathcal{D}_h. \qquad [264]$$

Combining this with [255] leads to

$$v_{m,n} \leq \frac{1}{4} \left(3 - (x_i^2 + y_j^2)\right) \| \varphi\|, \ \forall (x_m, y_n) \in \mathcal{D}_h, \qquad [265]$$

which implies using the fact that $x_i^2 + y_j^2 \geq 0$

$$v_{m,n} \leq \| \varphi\|, \ \forall (x_m, y_n) \in \mathcal{D}_h. \qquad [266]$$

Since $\bar{v}_h = -v_h$ satisfies, using [248]–[249]

$$\mathcal{L}_h \, \bar{v}_h(x_m, y_n) = \bar{\varphi}_{m,n}, \ , \ \forall (x_m, y_n) \in \Omega_h, \qquad [267]$$

where $\bar{\varphi}_{m,n} = -\varphi_{m,n}$,

$$\bar{v}_{m,n} = 0, \ \forall \, (x_m, y_n) \in \mathcal{D}_h \diagdown \Omega_h, \qquad [268]$$

and

$$\mathcal{L}_h \, \mathcal{P}_h(x_m, y_n) = \|\bar{\varphi}\|, \ \forall (x_m, y_n) \in \Omega_h, \qquad [269]$$

therefore, the previous reasoning, which allowed us to get [266], allows us to obtain

$$\bar{v}_{m,n} \leq \| \bar{\varphi}\|, \ \forall (x_m, y_n) \in \mathcal{D}_h. \qquad [270]$$

Which is equivalent to, since $\| \bar{\varphi}\| = \| \varphi\|$

$$- v_{m,n} \leq \| \varphi\|, \ \forall (x_m, y_n) \in \mathcal{D}_h. \qquad [271]$$

This with [266] implies

$$| v_{m,n}| \leq \| \varphi\|, \ \forall (x_m, y_n) \in \mathcal{D}_h, \qquad [272]$$

and therefore the stability of $\mathcal{L}_h$ is proved.

3. Consistency When applying $\mathcal{L}_h$, defined by [239], on the exact solution of [223]–[224], and using [64] and [65], we get, for all $(x_m, y_n) \in \Omega_h$ and $\bar{u}_h = (\, u(x_m, y_n))_{(x_m,y_n)\in\mathcal{D}_h}$

$$\begin{aligned} \mathcal{L}_h \bar{u}_h &= -\frac{\partial^2 u}{\partial x^2}(x_m, y_n) - \frac{\partial^2 u}{\partial y^2}(x_m, y_n) + \varepsilon_{m,n} \\ &= -\Delta\, u(x_m, y_n) + \varepsilon_{m,n} \\ &= f(x_m, y_n) + \varepsilon_{m,n}, \ \forall (x_m, y_n) \in \Omega_h \end{aligned} \quad [273]$$

where

$$|\varepsilon_{m,n}| \le \frac{h^2}{24}\{\max_{[0,1]^2} |\frac{\partial^4 u}{\partial x^4}(x, y)| + \max_{[0,1]^2} |\frac{\partial^4 u}{\partial y^4}(x, y)|\}, \ \forall (x_m, y_n) \in \Omega_h. \quad [274]$$

4. Convergence : we use now the two previous items of stability and consistency to prove the convergence of [227]–[231] towards the exact solution of [223]–[224]. To this end, we will use Theorem 6.2.

   Substracting [227] from [223] and using [273] to get

$$(\mathcal{L}_h(\bar{u}_h - u_h))_{m,n} = \varepsilon_{m,n}, \ \forall (x_m, y_n) \in \Omega_h, \quad [275]$$

   with

$$(\bar{u}_h - u_h)_{m,n} = 0, \ \forall (x_m, y_n) \in \mathcal{D}_h \diagdown \Omega_h. \quad [276]$$

   Applying Theorem 6.2 to get

$$\max_{(\mathbf{x}_m, y_n)\in\mathcal{D}_h} |\, u(x_m, y_n) - u_{m,n}| \le \frac{h^2}{24}\{\max_{[0,1]^2} |\frac{\partial^4 u}{\partial x^4}(x, y)| + \max_{[0,1]^2} |\frac{\partial^4 u}{\partial y^4}(x, y)|\}. \quad [277]$$

9 Finite difference methods for evolutive equations

So far we considered the finite difference approximation for stationary equations (do not depend on the time). We consider in this section evolutive equations (depend on the time).

9.1 A first example

Let us consider the following example of Cauchy problems:

$$u_t(x, t) - u_x(x, t) = \varphi(x, t), \ x \in \mathbb{R}, \ t \in [0, T], \quad [278]$$

and

$$u(x, 0) = \psi(x), \ x \in \mathbb{R}. \quad [279]$$

The numerical resolution of problem [278]–[279] can be performed via the following steps:

1. *Definition of the mesh*: since we have two variables x and t, we have then to define two discretization. The first one is performed on x–direction and the second one is performed in t–direction. The global mesh then is the *product* of these two discretizations. We denote then the global discretization by $\mathcal{V}$, where h and τ are two positive parameters

$$\mathcal{D} = \{(mh, n\tau), \ (m, n) \in \mathbb{Z} \times \{0, \ldots, N\}\}, \quad [280]$$

where $N \in \mathbb{N}$ satisfies $N\tau = T$.

2. *Finite difference scheme* Find $\{u_m^n;\ m \in \mathbb{Z},\ n = 1, \ldots, N\}$ such that

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \varphi(mh, n\tau),\ m \in \mathbb{Z},\ n = 0, \ldots, N-1, \tag{281}$$

with

$$u_m^0 = \psi(mh),\ m \in \mathbb{Z}. \tag{282}$$

To prove the well posedness of [281]–[282] as well as the convergence order of the solution of [281]–[282], we apply Theorem 6.2. Let us consider the operator $\mathcal{L}_\mathcal{D}$

$$\mathcal{L}_\mathcal{D}\, v_\mathcal{D} = \left(\frac{v_m^{n+1} - v_m^n}{\tau} - \frac{v_{m+1}^n - v_m^n}{h}\right)_{m\in\mathbb{Z},\ n=0,\ldots,N-1}, \tag{283}$$

with

$$v_m^0 = 0,\ m \in \mathbb{Z}, \tag{284}$$

and $v_\mathcal{D}$ is given by

$$v_\mathcal{D} = (\,u_m^n)_{m\in\mathbb{Z},\ n=0,\ldots,N}\,. \tag{285}$$

We will then verify the stability and consistency.

- *Stability* Let

$$\mathcal{L}_\mathcal{D}\, v_\mathcal{D} = \varphi_\mathcal{D}, \tag{286}$$

where

$$\varphi_\mathcal{D} = (\,\varphi_m^n)_{m\in\mathbb{Z},\ n=0,\ldots,N-1}\,. \tag{287}$$

Equation [286] is equivalent to

$$\frac{v_m^{n+1} - v_m^n}{\tau} - \frac{v_{m+1}^n - v_m^n}{h} = \varphi_m^n,\ \forall\,(m,n) \in \mathbb{Z} \times \{0, \ldots, N-1\}. \tag{288}$$

Which gives

$$v_m^{n+1} = \left(1 - \frac{\tau}{h}\right) v_m^n + \frac{\tau}{h} v_{m+1}^n + \tau\varphi_m^n,\ \forall\,(m,n) \in \mathbb{Z} \times \{0, \ldots, N-1\}. \tag{289}$$

We can see that we put $n = 0$ in the previous equation, we can compute v_m^1 for all $m \in \mathbb{Z}$ by using condition [284]. Therefore, successively on n we compute v_m^n for all $m \in \mathbb{Z}$. Which means that $\mathcal{L}_\mathcal{D}$ defined by [283] is invertible.

We assume the following assumption on the discretization $\mathcal{D}$ to get the stability of $\mathcal{L}_\mathcal{D}$.

Assumption 9.1 (An assumption on the ratio of space and time discretizations) We assume that the mesh (discretization) $\mathcal{D}$, given by [280], satisfies

$$\frac{\tau}{h} \leq 1. \tag{290}$$

Using then Assumption 9.1 (which means that $1-\frac{\tau}{h}\leq 0$) yields that

$$\begin{aligned} |\,v_m^{n+1}| &\leq \left(1-\frac{\tau}{h}+1\right)\max\left(|\,v_m^n|,|\,v_{m+1}^n|\right)+\tau|\varphi_m^n| \\ &\leq \sup_{m\in\mathbb{Z}}|\,v_m^n|+\tau\max_{m\in\mathbb{Z}}|\varphi_m^n|,\ \forall\,(m,n)\in\mathbb{Z}\times\{0,\ldots,N-1\}. \end{aligned} \qquad [291]$$

This implies that, using [284] and the fact that $N\tau=T$

$$\begin{aligned} \sup_{m\in\mathbb{Z}}|\,v_m^{n+1}| &\leq \sup_{m\in\mathbb{Z}}|\,v_m^n|+\tau\sup_{m\in\mathbb{Z}}|\varphi_m^n| \\ &\leq \sup_{m\in\mathbb{Z}}|\,v_m^n|+\tau\,M \\ &\leq \sup_{m\in\mathbb{Z}}|\,v_m^0|+N\tau\,M \\ &\leq T\,M,,\ \forall\ ,n\in\{0,\ldots,N-1\}. \end{aligned} \qquad [292]$$

where we have denoted $M=\sup_{n,m}|\,\varphi_m^n|$.

This yields that

$$\sup_{(m,n)}|\,v_m^n|\leq T\sup_{n,m}|\,\varphi_m^n|. \qquad [293]$$

- *Consitency* Let u be the solution of [278]–[279] and $u_{\mathcal{D}}=\left(u_m^n\right)_{m,n}$ be the finite difference solution of [281]–[282]. Let us denote by $\bar{u}_{\mathcal{D}}=\left(u(x_m,t_n)\right)_{m,n}$. Applying $\mathcal{L}_{\mathcal{D}}$ on $\bar{u}_{\mathcal{D}}-u_{\mathcal{D}}$ leads to, using equations [278] and [281], and Taylor expansion

$$\begin{aligned} \left(\mathcal{L}_{\mathcal{D}}(\bar{u}_{\mathcal{D}}-u_{\mathcal{D}})\right)_{m,n} &= \frac{u(x_m,t_{n+1})-u(x_m,t_n)}{\tau}-\frac{u(x_{m+1},t_n)-u(x_m,t_n)}{h}-\varphi(x_m,t_n) \\ &= u_t(x_m,t_n)-u_x(x_m,t_n)-\varphi(x_m,t_n)+\varepsilon_m^n \\ &= \varepsilon_m^n,\ \forall\,(m,n)\in\mathbb{Z}\times\{0,\ldots,N-1\}, \end{aligned} \qquad [294]$$

where, for all $(m,n)\in\mathbb{Z}\times\{0,\ldots,N-1\}$

$$\begin{aligned} |\varepsilon_m^n| &\leq \frac{1}{2}\left(h\sup_{\mathbb{R}\times[0,T]}|\,u_{tt}(x,t)|+\tau\sup_{\mathbb{R}\times[0,T]}|\,u_{xx}(x,t)|\right) \\ &\leq \frac{1}{2}\max\left(\sup_{\mathbb{R}\times[0,T]}|\,u_{tt}(x,t)|,\sup_{\mathbb{R}\times[0,T]}|\,u_{xx}(x,t)|\right)(\,h+\tau\,). \end{aligned} \qquad [295]$$

Therefore, [294] implies

$$\mathcal{L}_{\mathcal{D}}(\bar{u}_{\mathcal{D}}-u_{\mathcal{D}})=\varepsilon_{\mathcal{D}}, \qquad [296]$$

where $\varepsilon_{\mathcal{D}}=\left(\varepsilon_m^n\right)_{m,n}$.

Inequality [295] implies that, for all $(m,n)\in\mathbb{Z}\times\{0,\ldots,N-1\}$

$$|\varepsilon_m^n|\leq\frac{1}{2}\max\left(\sup_{\mathbb{R}\times[0,T]}|\,u_{tt}(x,t)|,\sup_{\mathbb{R}\times[0,T]}|\,u_{xx}(x,t)|\right)(\,h+\tau\,). \qquad [297]$$

This with [294] and [293] leads to

$$\sup_{(m,n)}|\,u(x_m,t_n)-u_m^n|\leq\frac{T}{2}\max\left(\sup_{\mathbb{R}\times[0,T]}|\,u_{tt}(x,t)|,\sup_{\mathbb{R}\times[0,T]}|\,u_{xx}(x,t)|\right)(\,h+\tau\,). \qquad [298]$$

Remark 7 (Convergence rate and uniqueness) The convergence order of the numerical scheme [281]–[282] given by estimate [298] implies the uniqueness of the solution of [278]–[279] in the sense that if u_1 and u_2 two *smooth* solutions for [278]–[279], we will have $\lim_{h\to 0,\tau\to 0} u_1(x_m,t_n) = \lim_{h\to 0,\tau\to 0} u_2(x_m,t_n)$, for all (n,m); indeed the triangle inequality and estimate [298] yields

$$\begin{aligned}
\sup_{(m,n)} |\,u_1(x_m,t_n)-u_2(x_m,t_n)| &\le \sup_{(m,n)} |\,u_1(x_m,t_n)-u_m^n| + \sup_{(m,n)} |\,u_m^n - u_2(x_m,t_n)| \\
&\le C\ (h+\tau), \qquad [299]
\end{aligned}$$

where

$$\begin{aligned}
C &= \frac{T}{2}\max\left(\sup_{\mathbb{R}\times[0,T]} |\,(u_1)_{tt}(x,t)|, \sup_{\mathbb{R}\times[0,T]} |\,(u_1)_{xx}(x,t)|\right) \\
&+ \frac{T}{2}\max\left(\sup_{\mathbb{R}\times[0,T]} |\,(u_2)_{tt}(x,t)|, \sup_{\mathbb{R}\times[0,T]} |\,(u_2)_{xx}(x,t)|\right). \qquad [300]
\end{aligned}$$

Tending h and τ to 0 in inequality [299] leads to, provided that second derivatives of u_1 and u_2 with respect to t and x are bounded

$$\lim_{h\to 0,\tau\to 0} \sup_{(m,n)} |\,u_1(x_m,t_n)-u_2(x_m,t_n)| = 0,\ \forall (m,n)\in\mathbb{Z}\times\{0,\ldots,N\}. \qquad [301]$$

what about if assumption 9.1 does not hold? In present section, we have proved that the finite difference solution [281]–[282] converges to the solution of the evolutive equation [278]–[279] in the sense of [298] provided that the following assumptions hold:

- assumption on u
  - $u \in \mathcal{C}^2(\mathbb{R}\times[0,T],\mathbb{R})$,
  - u_{tt} and u_{xx} are bounded over $\mathbb{R}\times[0,T]$.
- assumption on the mesh: the mesh (discretization) $\mathcal{D}$, given by [280], satisfies Assumption 9.1.

The following question deserves to be asked: what about if Assumption 9.1 does not hold?. We assume the following assumption on h and τ:

Assumption 9.2 (Relation between x and t discretizations) We assume that there exists a constant ξ, independent of h and τ such that:

$$\frac{\tau}{h} = \xi, \qquad [302]$$

where h (resp. τ) is the mesh step in the x (resp. t) discretization.

If Assumption 9.1 does not hold, i.e., $\frac{\tau}{h} > 1$, we will prove, under Assumption 9.2, that there is no convergence.

It seems that the consistency [294]–[295] remains hold, but we will prove that the convergences no longer holds in general (This implies that the stability does not hold.)

Equations [281]–[282] and Assumption 9.2 imply that, with $\dfrac{\tau}{h}=\xi$, $\bar{\zeta}=1-\xi$

$$u_m^{n+1}=\bar{\xi}\,u_m^n+\xi\,u_{m+1}^n+\tau\varphi_m^n,\ \forall\,(m,n)\in\mathbb{Z}\times\{0,\ldots,N-1\}. \qquad [303]$$

Putting $n=N-1$ and $m=0$ in the previous equation yields that

$$\begin{aligned}
u_0^N &= \bar{\xi}\,u_0^{N-1}+\xi\,u_1^{N-1}+\tau\varphi_0^{N-1}\\
&= \bar{\xi}\left(\bar{\xi}\,u_0^{N-2}+\xi\,u_1^{N-2}+\tau\varphi_0^{N-2}\right)+\xi\,\left(\bar{\xi}\,u_1^{N-2}+\xi\,u_2^{N-2}+\tau\varphi_1^{N-2}\right)+\tau\varphi_0^{N-1}\\
&= \bar{\xi}^2\,u_0^{N-2}+2\xi\,\bar{\xi}u_1^{N-2}+\xi^2\,u_2^{N-2}+\tau\varphi_0^{N-1}+\tau\left(\bar{\xi}\varphi_0^{N-2}+\xi\,\varphi_1^{N-2}\right)\\
&= \sum_{j=0}^{N}C_N^j\xi^j\bar{\xi}^{N-j}\ \psi\left(x_j\right)+\tau\varphi_0^{N-1}\\
&+ \tau\sum_{j=0}^{1}C_1^j\xi^j\bar{\xi}^{1-j}\varphi_j^{N-2}+\ldots+\tau\sum_{j=0}^{N-1}C_{N-1}^j\xi^j\bar{\xi}^{N-1-j}\varphi_j^0,
\end{aligned} \qquad [304]$$

where C_N^j is given by

$$C_N^j=\frac{N!}{j!(N-j)!}. \qquad [305]$$

We consider the case $\varphi(x,t)=0$, for all $(x,t)\in\mathbb{Z}\times[0,T]$. In addition to this, we assume that the function ψ satisfies

$$\psi(x)=1,\ \forall\,x\in[0,Nh]. \qquad [306]$$

Since $N\tau=T$, the previous choice for ψ could be written as

$$\psi(x)=1,\ \forall\,x\in[0,\frac{hT}{\tau}]. \qquad [307]$$

If Assumption 9.1 does not hold, then [290] no longer holds. This means that, using Assumption 9.2, $\alpha=\frac{Th}{\tau}=\frac{T}{\xi}<T$. Definition [307] becomes as

$$\psi(x)=1,\ \forall\,x\in[0,\alpha], \qquad [308]$$

where α is a positive constant only depending on T and the constant ξ of Assumption 9.2 and satisfies $0<\alpha<T$.

Since $\{jh;\ j=0,\ldots,N\}\subset\ [0,\alpha]$, one could deduce from [304] and [308] that (recall that $\dfrac{\tau}{h}=\xi$, $\bar{\xi}=1-\xi$.)

$$\begin{aligned}
u_0^N &= \sum_{j=0}^{N}C_N^j\xi^j\bar{\xi}^{N-j}\\
&= \left(\xi+\bar{\xi}\right)^N\\
&= 1.
\end{aligned} \qquad [309]$$

Let us remark that the solution of [278]–[279] is $u(x,t)=\psi(x+t)$, for all $(x,t)\in\mathbb{Z}\times[0,T]$. This implies that $u(0,T)=\psi(T)$.

The function $\psi(x)$ is already defined on $x \in [0, \alpha]$, by [308]. We define now the function $\psi(x)$ for $x \in [\alpha, +\infty]$. We consider the following choice

$$\psi(x) = 1 + e^{-\frac{1}{x^2-\alpha^2}}, \ \forall\, x \in (\alpha, +\infty]. \tag{310}$$

We can prove that $\psi \in \mathcal{C}^\infty[0, +\infty)$.
The choice [310] implies that $\psi(T) = 1 + e^{-\frac{1}{T-\alpha^2}}$; this with u_0^N computed in [309] yields that, since $u(0, T) = \psi(T)$

$$|u(0, T) - u_0^N| = e^{-\frac{1}{T-\alpha^2}}. \tag{311}$$

Which implies that

$$|u(0, T) - u_0^N| \not\to 0. \tag{312}$$

A direct proof of the no stability. We have proven, under Assumption 9.2, that there is no convergence when $\frac{\tau}{h} > 1$. This implies that, since we always have the consistency [296]–[297], that there is no stability for the operator $\mathcal{L}_\mathcal{D}$, defined by [283]–[285], when Assumption 9.2 holds and $\frac{\tau}{h} > 1$.
It is useful to show this non stability directy using the definition of stability.
Assume that $\mathcal{L}_\mathcal{D}$, defined by [283]–[285], is not stable. This implies that, there exists a constant C independent of h and τ such that:

$$\max_{n \in \{0,\ldots,N\}} \sup_{m \in \mathbb{Z}} |\, v_m^n| \le C \max_{n \in \{0,\ldots,N-1\}} \sup_{m \in \mathbb{Z}} |\, \varphi_m^n|, \tag{313}$$

where

$$\frac{v_m^{n+1} - v_m^n}{\tau} - \frac{v_{m+1}^n - v_m^n}{h} = \varphi_m^n, \ \forall\, (m, n) \in \mathbb{Z} \times \{0, \ldots, N-1\}, \tag{314}$$

and

$$v_m^0 = 0, \ m \in \mathbb{Z}. \tag{315}$$

Using the computation [304] yields that, for $(m, n) \in \mathbb{Z} \times \{1, \ldots, N\}$

$$v_m^n = \tau\varphi_0^{n-1} + \tau \sum_{j=0}^{1} C_1^j \xi^j \bar{\xi}^{1-j} \varphi_j^{n-2} + \ldots + \tau \sum_{j=0}^{n-1} C_{n-1}^j \xi^j \bar{\xi}^{n-1-j} \varphi_j^0, \tag{316}$$

where $\xi = \dfrac{\tau}{h}$ and $\bar{\xi} = 1 - \xi$ and $\xi > 1$.
Let us consider the choice

$$\varphi_m^n = (-1)^m, \forall\ (m, n) \in \mathbb{Z} \times \{0, \ldots, N-1\}. \tag{317}$$

With this choice, [316] becomes as

$$\begin{aligned} v_m^n &= \tau\left(1 + (1-2\xi)^1 + \ldots + (1-2\xi)^{n-1}\right) \\ &= \tau\frac{(1-2\xi)^n - 1}{(1-2\xi) - 1} \\ &= -\frac{\tau}{2\xi}\left((1-2\xi)^n - 1\right) \end{aligned} \tag{318}$$

The stability [313] could be written, since $|\varphi_m^n| = 1$ then as

$$\begin{aligned} \max_{n\in\{0,\ldots,N\}} \sup_{m\in\mathbb{Z}} |v_m^n| &\leq C \max_{n\in\{0,\ldots,N-1\}} \sup_{m\in\mathbb{Z}} |\varphi_m^n| \\ &\leq C, \end{aligned} \tag{319}$$

which implies that

$$|v_m^N| \leq C, \tag{320}$$

and then

$$\left|\frac{\tau}{2\xi}\left((1-2\xi)^N - 1\right)\right| \leq C. \tag{321}$$

On the other hand, since $1 - 2\xi < -1$ (and then $2\xi - 1 > 1$)

$$\begin{aligned} \lim_{\tau\to 0} \left|\frac{\tau}{2\xi}\left((1-2\xi)^N - 1\right)\right| &= \lim_{\tau\to 0} \left|\frac{\tau}{2\xi}(1-2\xi)^N\right| \\ &= \lim_{\tau\to 0} \frac{\tau}{2\xi}(2\xi - 1)^{\frac{T}{\tau}} \\ &= +\infty, \end{aligned} \tag{322}$$

which is contradiction with [321].

Remark 8 (No stability and no convergence) The previous result of the no stability does not imply the convergence. In fact, stabily with consistency imply the convergence in the sense of Theorem 6.2. The inverse of this previous statement is not true, i.e. the convergence does not imply neither the stability nor the consistency.

9.2 Second example

Let us consider the 1D example (This example took from my reply for a referee's report on a CRAS note.):

$$u_t(x,t) - u_{xx}(x,t) = 0,\ x \in \Omega = (0,1),\ t \in (0,1), \tag{323}$$

$$u(0,t) = u(1,t) = 0,\ t \in (0,1), \tag{324}$$

$$u(x,0) = \sin \pi x. \tag{325}$$

So, the analytical solution of [323]–[325] is

$$u(x,t) = \exp(-\pi^2 t) \sin \pi x. \tag{326}$$

We consider as a particular the meshes are uniform with $h = \frac{1}{M}$ (resp. $k = \frac{1}{N}$) in the space (resp. time) discretization, where M (resp. N) is integer. So, we consider the following scheme:

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} = 0,\ i \in [\![1, M-1]\!],\ n \in [\![0, N-1]\!], \tag{327}$$

$$u_0^n = u_M^n = 0,\ n \in [\![0, N]\!], \tag{328}$$

$$u_i^0 = \sin \pi x_i,\ i \in [\![0, M]\!]. \tag{329}$$

We can check that the finite volume solution of [327]–[329] is defined by, see [**?**, Pages 229–230] (there is some typos in the formula of λ_k given in [**?**, Page 230]!)

$$u_i^n = \lambda^n \sin \pi\, x_i = \lambda^n \sin \frac{\pi\, i}{M},\ i \in [\![0, M]\!],\ n \in [\![0, N]\!], \tag{330}$$

where

$$\lambda = \frac{1}{1 + 4r\sin^2 \frac{\pi}{2M}}, \tag{331}$$

with

$$r = \frac{k}{h^2}. \tag{332}$$

Let us examine now the convergence order, we will use mainly Taylor's expansions.

1. finite difference convergence order, i.e. $\mathbb{L}^\infty(\mathbb{L}^\infty)$.

   (a) First method: *stability* and *consistency*: The convergence in the finite difference methods can be obtained, as usual, as the product of the *stability* and the *consistency*. The convergence order in finite difference methods can be obtained via the order of the approximation of the operator $u_t - u_{xx}$. Indeed, let u be the solution of [323]–[325]. We have

   $$\begin{aligned}&\frac{u(x_i, t_{n+1}) - u(x_i, t_n)}{k} - \frac{u(x_{i+1}, t_n) - 2u(x_i, t_n) + u(x_{i-1}, t_n)}{h^2}\\ = \quad & u_t(x_i, t_n) - u_{xx}(x_i, t_n) + 0(k + h^2) = 0(k + h^2).\end{aligned} \tag{333}$$

   So the convergence order in $\mathbb{L}^\infty(\mathbb{L}^\infty)$ is $k + h^2$

   (b) Second method: Direct method. The subject of this method is to compute the error $u(x_i, t_n) - u_i^n$. We first compute an expansion for u_i^n given by [330] and then we coompute the difference between this expansion and the expression of $u(x_i, t_n)$ defined by replacing (x, t) in [326] by (x_i, t_n)

   $$\sin x = x + O(x^3). \tag{334}$$

   which gives

   $$\sin^2 x = x^2 + O(x^4). \tag{335}$$

   so

   $$\sin^2 \frac{\pi}{2} h = \frac{\pi^2}{4} h^2 + O(h^4). \tag{336}$$

   Which gives, since $h = \frac{1}{M}$

   $$\begin{aligned}1 + 4r\sin^2 \frac{\pi}{2M} &= 1 + 4\frac{k}{h^2}\left(\frac{\pi^2}{4} h^2 + O(h^4)\right)\\ &= 1 + k\pi^2 + O(h^2 k).\end{aligned} \tag{337}$$

   So

   $$\begin{aligned}\lambda^n &= \exp(n \log \lambda) = \exp(-\frac{t_n}{k} \log(1 + 4r\sin^2 \frac{\pi}{2M}))\\ &= \exp(-\frac{t_n}{k} \log(1 + k\pi^2 + O(h^2 k))).\end{aligned} \tag{338}$$

But

$$\begin{aligned}\log(1+k\pi^2+O(h^2k)) &= k\pi^2+O(h^2k)+O\left(\left(k+h^2k\right)^2\right)\\ &= k\pi^2+O(h^2k)+O\left((h^2+1)^2k^2\right).\end{aligned} \quad [339]$$

Multiplying the previous expansion by $-\frac{t_n}{k}$, we get, since $t_n \in [0,1]$ and $0 < h \leq 1$ (so h^2+1 is bounded)

$$-\frac{t_n}{k}\log(1+k\pi^2+O(h^2k)) = -\pi^2 t_n + O(k+h^2), \quad [340]$$

so, since $\exp(x) = 1 + 0(x)$ and $\exp\left(-\pi^2 t_n\right) \leq 1$

$$\begin{aligned}\exp\left(-\frac{t_n}{k}\log(1+k\pi^2+O(h^2k))\right) &= \exp\left(-\pi^2 t_n\right)\,\exp\left(O(k+h^2)\right)\\ &= \exp\left(-\pi^2 t_n\right)\{1+O(k+h^2)\}\\ &= \exp\left(-\pi^2 t_n\right)+O(k+h^2).\end{aligned} \quad [341]$$

Replacing this in [338], we get

$$\lambda^n = \exp\left(-\pi^2 t_n\right) + O(k+h^2). \quad [342]$$

Inserting this in the expression of u_i^n given by [330], using the fact that $\sin\dfrac{\pi\, i}{M} \leq 1$, using the expression of u given by [326], we get

$$\begin{aligned}u_i^n &= \left(\exp\left(-\pi^2 t_n\right)+O(k+h^2)\right)\sin \pi\, x_i\\ &= \exp\left(-\pi^2 t_n\right)\sin \pi\, x_i + O(k+h^2)\\ &= u(x_i,t_n)+O(k+h^2),\ i \in [\![0,M]\!],\ n \in [\![0,N]\!],\end{aligned} \quad [343]$$

which implies that

$$\max_{(i,n)\in[\![0,M]\!]\times[\![0,N]\!]} |\, u_i^n - u(x_i,t_n)| \leq C(k+h^2). \quad [344]$$

2. Convergence order in $\mathbb{L}^\infty(\mathcal{W}^{1,\infty})$: using the expressions [330] , [326] and [342], we get

$$\begin{aligned}\frac{u_{i+1}^n-u_i^n}{h} &- u_x(x_i,t_n) = \frac{\lambda^n}{h}\left(\sin(\pi\, x_i)\left(\cos(\pi\, h)-1\right)+\cos(\pi\, x_i)\sin(\pi\, h)\right)\\ &- \pi\exp\left(-\pi^2 t_n\right)\cos(\pi\, x_i)\\ &= \left(\exp\left(-\pi^2 t_n\right)+O(k+h^2)\right) r_i^h - \pi\exp\left(-\pi^2 t_n\right)\cos(\pi\, x_i)\\ &= \exp\left(-\pi^2 t_n\right)\left(r_i^h-\pi\cos(\pi\, x_i)\right)+O(k+h^2)r_i^h\end{aligned} \quad [345]$$

where

$$r_i^h = \frac{\sin(\pi\, x_i)\left(\cos(\pi\, h)-1\right)+\cos(\pi\, x_i)\sin(\pi\, h)}{h}. \quad [346]$$

Using the triangle inequality and the fact that $\exp\left(-\pi^2 t_n\right) \leq 1$, [345] implies

$$|\frac{u_{i+1}^n-u_i^n}{h} - u_x(x_i,t_n)| \leq |\, r_i^h - \pi\cos(\pi\, x_i)| + O(k+h^2)|\, r_i^h|. \quad [347]$$

(a) *Expansion for* $r_i^h - \pi\cos(\pi\, x_i)$. We have, since $|\sin(x)|, |\cos(x)| \leq 1$

$$
\begin{aligned}
|\, r_i^h - \pi\cos(\pi\, x_i)| &= \left|\frac{\sin(\pi\, x_i)\,(\cos(\pi\, h) - 1) + \cos(\pi\, x_i)\sin(\pi\, h)}{h} - \pi\cos(\pi\, x_i)\right| \\
&\leq \frac{|\cos(\pi\, h) - 1|}{h} + \left|\frac{\sin(\pi\, h)}{h} - \pi\right| \\
&\leq O(h) + O(h^2) \\
&\leq O(h) \qquad [348]
\end{aligned}
$$

(b) *Estimate for* r_i^h. Estimate [348] implies that r_i^h is bounded, i.e. for some positive constant C independent of h and k such that

$$\max_{(i,n)\in[\![0,M]\!]\times[\![0,N]\!]} |\, r_i^h| \leq C. \qquad [349]$$

So [347] with [348] and [349] implies, for some positive constant C independent of h and k such that

$$\max_{(i,n)[\![0,M]\!]\times[\![0,N]\!]} |\frac{u_{i+1}^n - u_i^n}{h} - u_x(x_i, t_n)| \leq C(k+h). \qquad [350]$$

3. Convergence order in $\mathcal{W}^\infty(\mathbb{L}^{1,\infty})$: using the expressions [330] , [326] and [342], using a similar reasoning to that used in [338]–[342], we get

$$
\begin{aligned}
|\frac{u_i^{n+1} - u_i^n}{k} &- u_t(x_i, t_n)| = |-\frac{4}{h^2}\lambda^{n+1}\sin(\pi\, x_i)\sin^2(\frac{\pi}{2}h) + \pi^2\exp\left(-\pi^2 t_n\right)\sin(\pi\, x_i)| \\
&= |\sin(\pi\, x_i)\left(-\frac{4}{h^2}\lambda^{n+1}\sin^2(\frac{\pi}{2}h) + \pi^2\exp\left(-\pi^2 t_n\right)\right)| \\
&\leq |-\frac{4}{h^2}\lambda^{n+1}\sin^2(\frac{\pi}{2}h) + \pi^2\exp\left(-\pi^2 t_n\right)|. \qquad [351]
\end{aligned}
$$

Let us first simplify the expansion $-\frac{4}{h^2}\lambda^{n+1}\sin^2(\frac{\pi}{2}h)$ on the r.h.s. of the previous inequality and then replace it by the result

$$
\begin{aligned}
-\frac{4}{h^2}\lambda^{n+1}\sin^2(\frac{\pi}{2}h) &= -\frac{4}{h^2}\left(\exp\left(-\pi^2 t_{n+1}\right) + 0(k+h^2)\right)\sin^2(\frac{\pi}{2}h) \\
&= -\frac{4}{h^2}\exp\left(-\pi^2 t_{n+1}\right)\sin^2(\frac{\pi}{2}h) + +\frac{4}{h^2}0(k+h^2)\sin^2(\frac{\pi}{2}h) \\
&= -\frac{4}{h^2}\exp\left(-\pi^2 t_{n+1}\right)\sin^2(\frac{\pi}{2}h) + +0(k+h^2) \\
&= -\frac{4}{h^2}\exp\left(-\pi^2 t_n\right)\exp\left(-\pi^2 h\right)\sin^2(\frac{\pi}{2}h) + 0(k+h^2) \\
&= -\frac{4}{h^2}\exp\left(-\pi^2 t_n\right)(1 + O(h))\left(\frac{\pi^2}{4}h^2 + O(h^4)\right) + 0(k+h^2) \\
&= -\frac{4}{h^2}\exp\left(-\pi^2 t_n\right)\left(\frac{\pi^2}{4}h^2 + O(h^4)\right) + 0(k+h^2) \\
&= -\pi^2\exp\left(-\pi^2 t_n\right) + 0(k+h^2). \qquad [352]
\end{aligned}
$$

Inserting this in r.h.s. of [351], we get, for some positive constant C independent of h and k such that

$$\max_{(i,n)\in[\![0,M]\!]\times[\![0,N]\!]} |\frac{u_i^{n+1} - u_i^n}{k} - u_t(x_i, t_n)| \leq C(k+h^2). \qquad [353]$$

Some numerical tests: The present pragraph is devoted to report some numerical tests justifying that [344] ($\mathbb{L}^\infty(\mathbb{L}^\infty)$–estimate), [350] ($\mathbb{L}^\infty(\mathcal{W}^{1,\infty})$–estimate), [353] ($\mathcal{W}^{1,\infty}(\mathbb{L}^\infty)$–estimate)

M	N	$\frac{\lvert e\rvert_{\mathbb{L}^\infty(\mathbb{L}^\infty)}}{k+h^2}$	$\frac{\lvert e\rvert_{\mathcal{W}^{1,\infty}(\mathbb{L}^\infty)}}{k+h^2}$	$\frac{\lvert e\rvert_{\mathbb{L}^\infty(\mathcal{W}^{1,\infty})}}{k+h}$
25	25	1.527701	68.149626	2.6430409
50	25	1.562728	69.916978	3.3336229
100	25	1.5729726	70.429576	3.955193
150	25	1.5749271	70.527375	4.2331084
200	25	1.5756177	70.561933	4.3900979
250	25	1.5759389	70.578005	4.4909097
300	25	1.5761139	70.586761	4.5611038
350	25	1.5762196	70.59205	4.6127839
400	25	1.5762883	70.595488	4.652419
450	25	1.5763355	70.597848	4.6837797
450	50	1.6781826	81.595325	4.7417958
450	100	1.7428064	88.696505	4.4888874
450	150	1.7659023	91.369813	4.1848073
450	350	1.7918674	94.594441	3.2750937
450	400	1.7940874	94.896783	3.1193867
450	450	1.7957339	95.128787	2.985172
500	450	1.7963641	95.165563	3.1078235
500	500	1.7976785	95.352012	2.9875977
500	550	1.7987003	95.502067	2.8829679
500	600	1.799504	95.624664	2.7926277
500	650	1.8001355	95.726039	2.7486286

where $h = 1/(M+1)$ and $k = 1/(N+1)$, with M (resp. N) is the number of the spatial mesh points (resp. temporal mesh points) without $\{0,1\}$. So the finite volume scheme [327]–[329] leads to sets of systems which can be sloved successively starting from the level $n = 0$:

$$\mathcal{A}\,U^{n+1} = U^n,\ n \in [\![0, N]\!], \qquad [354]$$

with

$$U^0 = (\,\sin(\pi\,x_i))_1^M\,, \qquad [355]$$

and $\mathcal{A}$ is the $M \times M$ matrix

$$\mathcal{A} = \begin{pmatrix} 1+2r & -r & 0 & & \dots & 0 \\ -r & 1+2r & -r & 0 & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & & 0 & -r & 1+2r \end{pmatrix} \qquad [356]$$

10 Appendix: comments on the regularity of the exact solution required in the finite difference approximation

The useful information, I think, I'm quoting in this section are given in [GOD 77, Chapter 6, Pages 239–253].

In the previous sections, and in order to get the convergence, we assume that the exact solution is smooth. We assumed that the exact solution with its derivatives (or partial derivatives) are bounded. These regularity assumption can be replaced using Sobolev spaces instead of the classical spaces (Would say, the spaces of functions which with their derivatives or partial derivatives are bounded.) In fact, the finite difference method is based on the approximation of the derivatives, which appear in differential or partial differential equation under consideration, by differential quotients.

Some physical process, the functions (given data) are not even derivable. Indeed, in some evolutive process, the exact solutions have jumps even the initial data are smooth. These evolutive equations do not have regular (smooth) solutions. We need then to introduce another sense of the exact solution in which the discontinuous data are allowed. We have at least two issues:

1. we write the equations of conservation laws in some integral forms instead of the differential or partial differential forms. The integration of functions (even they have some points where they are not continue) included in these conservation laws is allowed and it has sense. These integral forms may be, for instance, weak formulations or entropic forms.

2.

10.1 Some examples

In this subsection, we quote some example explaing the use of the two previous issues.

1. first example: let us consider the following equation, for a given positive number $T > 0$ and a given function $\psi(x)$ defined on $x \in \mathbb{R}$

$$u_t(x,t) + u(x,t)u_x(x,t) = 0,\ \forall (x,t) \in \mathbb{R} \times (0,T), \quad [357]$$

with

$$u(x,0) = \psi(x),\ \forall x \in \mathbb{R}. \quad [358]$$

Equation [357]–[358] is the simplest model in *fluid mecanique*. It is also called, in some references, the Bürgers equation.

(a) the discontinuities points: we first assume that the exact solution u is smooth and let us consider the lines $x(t)$ defined via the following equation:

$$\frac{dx}{dt} = u(x,t). \quad [359]$$

The lines $x(t)$ called the *characteristics* of equation $u_t(x,t)+u(x,t)u_x(x,t)=0$. On these lines, $u(x,t)$ can be written as a function in t instead of (x,t). Indeed, on these lines $u(x,t)=u(x(t),t)=u(t)$; and then using an integrtion of composed functions, [359], and [357]

$$\begin{aligned}\frac{du}{dt}(x(t),t) &= \frac{\partial u}{\partial t}(x(t),t)+\frac{\partial u}{\partial x}\frac{dx}{dt}\\ &= \frac{\partial u}{\partial t}+u(x,t)\frac{\partial u}{\partial x}\\ &= 0. \qquad [360]\end{aligned}$$

Which implies that, there exists a constant still denoted by u

$$u(x(t),t)=u. \qquad [361]$$

This with [359] leads to

$$x(t)=ut+x_0. \qquad [362]$$

11 Programme

- Chapter one: Finite differences methods, see [SMI 85] and [GOD 77].
- Chapter two: Finite volume methods, [EYM 00].
- Chapter three: Finite element methods, [CIA 78].
- Chapter four: Spectrale methods, [BER 92]

References

[BER 92] C. Bernardi and Y. Maday: Approximation Spectrale de Problèmes aux limites. *Springer–Verlag France, Paris*, 1992.

[CIA 78] P. G. Ciarlet: The Finite Element Method for Elliptic Problems. *Norhh Holand, Amsterdam*, 1978.

[EYM 00] R. Eymard, T. Gallouët and R. Herbin: Finite volume methods. *Handbook of Numerical Analysis. P. G. Ciarlet and J. L. Lions (eds.)*, **VII** , 723-1020, 2000.

[GOD 77] S. Godunov and V. Riabenki: Schémas aux Differences.*Editions Mir, Moscow, (French)*, 1977.

[SMI 85] G. D. Smith: Numerical Solution of Partial Differential Equations: Finite Difference Methods. *Oxford University Press, Third edition*, 1985.