Lecture Notes, Course on Numerical Analysis

Guillaume Bal *

October 20, 2008

Contents

*Department of Applied Physics & Applied Mathematics, Columbia University, New York NY, 10027, gb2030@columbia.edu, http://www.columbia.edu/∼gb2030

1 Ordinary Differential Equations

An Ordinary Differential Equation is an equation of the form

$$\begin{aligned} &\frac{dX(t)}{dt} = f(t, X(t)), \qquad t \in (0, T), \\ &X(0) = X_0. \end{aligned} \tag{1.1}$$

We have existence and uniqueness of the solution when $f(t, x)$ is continuous and Lipschitz with respect to its second variable, i.e. there exists a constant C independent of t, x, y such that

$$|f(t, x) - f(t, y)| \leq C|x - y|. \tag{1.2}$$

If the constant is independent of time for $t > 0$, we can even take $T = +\infty$. Here X and $f(t, X(t))$ are vectors in $\mathbb{R}^n$ for $n \in \mathbb{N}$ and $|\cdot|$ is any norm in $\mathbb{R}^n$.

Remark 1.1 Throughout the text, we use the symbol "C" to denote an arbitrary constant. The "value" of C may change at every instance. For example, if $u(t)$ is a function of t bounded by 2, we will write

$$C|u(t)| \leq C, \tag{1.3}$$

even if the two constants "C" on the left- and right-hand sides of the inequality may be different. When the value "2" is important, we will replace the above right-hand side by $2C$. In our convention however, $2C$ is just another constant "C".

Higher-order differential equations can always been put in the form (1.1), which is quite general. For instance the famous *harmonic oscillator* is the solution to

$$x'' + \omega^2 x = 0, \qquad \omega^2 = \frac{k}{m}, \tag{1.4}$$

with given initial conditions $x(0)$ and $x'(0)$ and can be recast as

$$\frac{d}{dt}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(t) = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(t), \qquad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(0) = \begin{pmatrix} x(0) \\ x'(0) \end{pmatrix}. \tag{1.5}$$

Exercise 1.1 Show that the implied function f in (1.5) is Lipschitz.

The Lipschitz condition is quite important to obtain uniqueness of the solution. Take for instance the case $n = 1$ and the function $x(t) = t^2$. We easily obtain that

$$x'(t) = 2\sqrt{x(t)} \quad t \in (0, +\infty), \qquad x(0) = 0.$$

However, the solution $\tilde{x}(t) \equiv 0$ satisfies the same equation, which implies non-uniqueness of the solution. This remark is important in practice: when an equation admits several solutions, any sound numerical discretization is likely to pick one of them, but not necessarily the solution one is interested in.

Exercise 1.2 Show that the function $f(t)$ above is *not* Lipschitz.

How do we discretize (1.1)? There are two key ingredients in the derivation of a numerical scheme: first to remember the definition of a derivative

$$\frac{df}{dt}(t) = \lim_{h\to 0} \frac{f(t+h)-f(t)}{h} = \lim_{h\to 0} \frac{f(t)-f(t-h)}{h} = \lim_{h\to 0} \frac{f(t+h)-f(t-h)}{2h},$$

and second (which is quite related) to remember Taylor expansions to approximate a function in the vicinity of a given point:

$$f(t+h) = f(t) + hf'(t) + \frac{h^2}{2}f''(t) + \ldots + \frac{h^n}{n!}f^{(n)}(t) + \frac{h^{n+1}}{(n+1)!}f^{(n+1)}(t+s), \quad (1.6)$$

for $0 \le s \le t$ assuming the function f is sufficiently regular.

With this in mind, let us fix some $\Delta t > 0$, which corresponds to the size of the interval between successive points where we try to approximate the solution of (1.1). We then approximate the derivative by

$$\frac{dX}{dt}(t) \approx \frac{X(t+\Delta t) - X(t)}{\Delta t}. \quad (1.7)$$

It remains to approximate the right-hand side in (1.1). Since $X(t)$ is supposed to be known by the time we are interested in calculating $X(t+\Delta t)$, we can choose $f(t, X(t))$ for the right-hand side. This gives us the so-called *explicit Euler* scheme

$$\frac{X(t+\Delta t) - X(t)}{\Delta t} = f(t, X(t)). \quad (1.8)$$

Let us denote by $T_n = n\Delta t$ for $0 \le n \le N$ such that $N\Delta t = T$. We can recast (1.8) as

$$X_{n+1} = X_n + \Delta t f(n\Delta t, X_n), \qquad X_0 = X(0). \quad (1.9)$$

This is a fully discretized equation that can be solved on the computer since $f(t,x)$ is known.

Another choice for the right-hand side in (1.1) is to choose $f(t+\Delta t, X(t+\Delta t))$. This gives us the *implicit Euler* scheme

$$X_{n+1} = X_n + \Delta t f((n+1)\Delta t, X_{n+1}), \qquad X_0 = X(0). \quad (1.10)$$

Notice that this scheme necessitates to solve an equation at every step, namely to find the inverse of the function $x \mapsto x - f(n\Delta t, x)$. However we will see that this scheme is more stable than its explicit relative in many practical situations and is therefore often a better choice.

Exercise 1.3 Work out the explicit and implicit schemes in the case $f(t,x) = \lambda x$ for some real number $\lambda \in \mathbb{R}$. Write explicitly X_n for both schemes and compare with the exact solution $X(n\Delta t)$. Show that in the stable case ($\lambda < 0$), the explicit scheme can become unstable and that the implicit scheme cannot. Comment.

We now need to justify that the schemes we have considered are good approximations of the exact solution and understand how good (or bad) our approximation is. This requires to analyze the behavior of X_n as $\Delta t \to 0$, or equivalently the number of points

N to discretize a fixed interval of time $(0, T)$ tends to infinity. In many problems, the strategy to do so is the following. First we need to assume that the exact problem (1.1) is *properly posed*, i.e. roughly speaking has good stability properties: when one changes the initial conditions a little, one expects the solution not to change too much in the future. A second ingredient is to show that the scheme is *stable*, i.e. that our solution X_n remains bounded independently of n and Δt. The third ingredient is to show that our scheme is *consistent* with the equation (1.1), i.e. that locally it is a good approximation of the real equation.

Let us introduce the solution function $g(t, X) = g(t, X; \Delta T)$ defined by

$$g(t, X) = X(t + \Delta T), \tag{1.11}$$

where

$$\begin{array}{l} \dfrac{dX}{dt}(t+\tau) = f(t+\tau, X(t+\tau)), \qquad 0 \le \tau \le \Delta t \\ X(t) = X. \end{array} \tag{1.12}$$

This is nothing but the function that maps a solution of the ODE at time t to the solution at time $t + \Delta t$.

Similarly we introduce the approximate solution function $g_{\Delta t}(t, X)$ for $t = n\Delta t$ by

$$X_{n+1} = g_{\Delta t}(n\Delta t, X_n) \quad \text{ for } 0 \le n \le N - 1. \tag{1.13}$$

For instance, for the explicit Euler scheme, we have

$$g_{\Delta t}(n\Delta t, X_n) = X_n + \Delta t f(n\Delta t, X_n). \tag{1.14}$$

For the implicit Euler scheme, we have

$$g_{\Delta t}(n\Delta t, X_n) = (x - \Delta t f((n+1)\Delta t, x))^{-1}(X_n),$$

assuming that this inverse function is well-defined. Because the explicit scheme is much simpler to analyze, we assume that (1.14) holds from now on.

That the equation is *properly posed* (a.k.a. *well-posed*) in the sense of Hadamard means that

$$|g(t, X) - g(t, Y)| \le (1 + C\Delta t)|X - Y|, \tag{1.15}$$

where the constant C is independent of t, X and Y. The meaning of this inequality is the following: we do not want the difference $X(t + \Delta t) - Y(t + \Delta t)$ to be more than $(1 + C\Delta t)$ times what it was at time t (i.e. $|X - Y| = |X(t) - Y(t)|$) during the time interval Δt.

We then have to prove that our scheme is *stable*, that is to say

$$|X_n| \le C(1 + |X_0|), \tag{1.16}$$

where C is independent of X_0 and $0 \le n \le N$. Notice from the Lipschitz property of f (with $x = x$ and $y = 0$) and from (1.14) that

$$|X_{n+1}| = |g_{\Delta t}(n\Delta t, X_n)| \le (1 + C\Delta t)|X_n| + C\Delta t,$$

where C is a constant. This implies by induction that

$$|X_n| \leq (1 + C\Delta t)^n |X_0| + \sum_{k=0}^{n} (1 + C\Delta t)^k C\Delta t \leq (1 + C\Delta t)^N (N\Delta t C + |X_0|).$$

for $1 \leq n \leq N$. Now $N\Delta t = T$ and

$$(1 + \frac{CT}{N})^N \leq e^{CT},$$

which is bounded independently of Δt. This implies the stability of the explicit Euler scheme.

We then prove that the scheme is *consistent* and obtain the local error of discretization. In our case, it consists of showing that

$$|g(n\Delta t, X) - g_{\Delta t}(n\Delta t, X)| \leq C\Delta t^{m+1}(1 + |X|), \tag{1.17}$$

for some positive m, the order of the scheme. Notice that this is a *local* estimate. Assuming that we know the initial condition at time $n\Delta t$, we want to make sure that the exact and approximate solutions at time $(n + 1)\Delta t$ are separated by an amount at most of order Δt^2. Such a result is usually obtained by using Taylor expansions. Let us consider the explicit Euler scheme and assume that the solution $X(t)$ of the ODE is sufficiently smooth so that the solution with initial condition at $n\Delta t$ given by $X = X(n\Delta t)$ satisfies

$$X((n + 1)\Delta t) = X(n\Delta t) + \Delta t \dot{X}(n\Delta t) + \frac{\Delta t^2}{2} \ddot{X}(n\Delta t + s), \tag{1.18}$$

for some $0 \leq s \leq \Delta t$, where $|\ddot{X}(n\Delta t + s)| \leq C(1 + |X|)$. Notice that the above property is a regularity property of the equation, not of the discretization. *We prove convergence of the explicit Euler scheme only when the above regularity conditions are met.* These are sufficient conditions (though not necessary ones) to obtain convergence and can be shown rigorously when the function $f(t, x)$ is of class C^1 for instance. This implies that

$$\begin{aligned} X((n + 1)\Delta t) &= g(n\Delta t, X(n\Delta t)) \\ &= X(N\Delta t) + \Delta t f(n\Delta t, X(n\Delta t)) + \frac{\Delta t^2}{2} \ddot{X}(n\Delta t + s) \\ &= g_{\Delta t}(n\Delta t, X(n\Delta t)) + \frac{\Delta t^2}{2} \ddot{X}(n\Delta t + s). \end{aligned}$$

This implies (1.17) and the *consistency* of the scheme with $m = 1$ (the Euler explicit scheme is first-order). Another way of interpreting the above result is as follows: the exact solution $X((n + 1)\Delta t)$ locally solves the discretized equation (whose solution is $g_{\Delta t}(n\Delta t, X(n\Delta t))$ at time $(n + 1)\Delta t$) up to a small term, here of order Δt^2.

Exercise 1.4 Consider the harmonic oscillator (1.5). Work out the functions g and g_Δ for the explicit and implicit Euler schemes. Show that (1.15) is satisfied, that both schemes are stable (i.e. (1.16) is satisfied) and that both schemes are consistent and of order $m = 1$ (i.e. (1.17) is satisfied with $m = 1$).

Once we have well-posedness of the exact equation as well as stability and consistency of the scheme (for a scheme of order m), we obtain convergence as follows. We have

$$\begin{aligned}
&|X((n+1)\Delta t) - X_{n+1}| = |g(n\Delta t, X(n\Delta t)) - g_\Delta(n\Delta t, X_n)| \\
&\leq |g(n\Delta t, X(n\Delta t)) - g(n\Delta t, X_n)| + |g(n\Delta t, X_n) - g_{\Delta t}(n\Delta t, X_n)| \\
&\leq (1 + C\Delta t)|X(n\Delta t) - X_n| + C\Delta t^{m+1}(1 + |X_n|) \\
&\leq (1 + C\Delta t)|X(n\Delta t) - X_n| + C\Delta t^{m+1}.
\end{aligned}$$

The fist line is the definition of the operators. The second line uses the *triangle inequality*

$$|X + Y| \leq |X| + |Y|, \tag{1.19}$$

which holds for every norm by definition. The third line uses the well-posedness of the ODE and the consistency of the scheme. The last line uses the stability of the scheme. Recall that C is here the notation for a constant that may change every time the symbol is used.

Let us denote the error between the exact solution and the discretized solution by

$$\varepsilon_n = |X(n\Delta t) - X_n|. \tag{1.20}$$

The previous calculations have shown that for a scheme of order m, we have

$$\varepsilon_{n+1} \leq C\Delta t^{m+1} + (1 + C\Delta t)\varepsilon_n.$$

Since $\varepsilon_0 = 0$, we easily deduce from the above relation (for instance by induction) that

$$\varepsilon_n \leq C\Delta t^{m+1} \sum_{k=0}^{n} (1 + C\Delta t)^k \leq C\Delta t^{m+1} n,$$

since $(1 + C\Delta t)^k \leq C$ uniformly for $0 \leq k \leq N$. We deduce from the above relation that

$$\varepsilon_n \leq C\Delta t^m,$$

since $n \leq N = T\Delta t^{-1}$ This shows that the explicit Euler scheme is of order $m = 1$: for all intermediate time steps $n\Delta t$, the error between $X(n\Delta t)$ and X_n is bounded by a constant time Δt. Obviously, as Δt goes to 0, the discretized solution converges to the exact one.

Exercise 1.5 Program in Matlab the explicit and implicit Euler schemes for the harmonic oscillator (1.5). Find numerically the order of convergence of both schemes.

Higher-Order schemes. We have seen that the Euler scheme is first order accurate. For problems that need to be solved over long intervals of time, this might not be accurate enough. The obvious solution is to create a more accurate discretization. Here is one way of constructing a second-order scheme. To simplify the presentation we assume that $X \in \mathbb{R}$, i.e. we consider a scalar equation.

We first realize that m in (1.17) has to be 2 instead of 1. We also realize that the local approximation to the exact solution must be compatible to the right order with the Taylor expansion (1.18), which we push one step further to get

$$X((n+1)\Delta t) = X(n\Delta t) + \Delta t \dot{X}(n\Delta t) + \frac{\Delta t^2}{2}\ddot{X}(n\Delta t) + \frac{\Delta t^3}{6}\dddot{X}(n\Delta t + s) \qquad (1.21)$$

for some $0 \leq s \leq \Delta t$. Now from the derivation of the first-order scheme, we know that what we need is an approximation such that

$$X((n+1)\Delta t) = g^{(2)}_{\Delta t}(n\Delta t, X) + O(\Delta t^3).$$

An obvious choice is then to choose

$$g^{(2)}_{\Delta t}(n\Delta t, X) = X(n\Delta t) + \Delta t \dot{X}(n\Delta t) + \frac{\Delta t^2}{2}\ddot{X}(n\Delta t).$$

This is however not explicit since $\dot{X}$ and $\ddot{X}$ are not known yet (only the initial condition $X = X(n\Delta t)$ is known when we want to construct $g^{(2)}_{\Delta t}(n\Delta t, X)$). However, as previously, we can use the equation that $X(t)$ satisfies and get that

$$\begin{aligned} \dot{X}(n\Delta t) &= f(n\Delta t, X) \\ \ddot{X}(n\Delta t) &= \frac{\partial f}{\partial t}(n\Delta t, X) + \frac{\partial f}{\partial x}(n\Delta t, X) f(n\Delta t, X), \end{aligned}$$

by applying the chain rule to (1.1) at $t = n\Delta t$. Assuming that we know how to differentiate the function f, we can then obtain the second-order scheme by defining

$$\begin{aligned} g^{(2)}_{\Delta t}(n\Delta t, X) &= X + \Delta t f(n\Delta t, X) \\ &\quad + \frac{\Delta t^2}{2}\Big(\frac{\partial f}{\partial t}(n\Delta t, X) + \frac{\partial f}{\partial x}(n\Delta t, X) f(n\Delta t, X)\Big). \end{aligned} \qquad (1.22)$$

Clearly by construction, this scheme is consistent and second-order accurate. The regularity that we now need from the exact equation (1.1) is that $|\dddot{X}(n\Delta t + s)| \leq C(1 + |X|)$. We shall again assume that our equation is nice enough so that the above constraint holds (this can be shown when the function $f(t, x)$ is of class C^2 for instance).

Exercise 1.6 It remains to show that our scheme $g^{(2)}_{\Delta t}(n\Delta t, X)$ is stable. This is left as an exercise. *Hint:* show that $|g^{(2)}_{\Delta t}(n\Delta t, X)| \leq (1 + C\Delta t))|X| + C\Delta t$ for some constant C independent of $1 \leq n \leq N$ and X.

Exercise 1.7 Work out a second-order scheme for the harmonic oscillator (1.5) and show that all the above constraints are satisfied.

Exercise 1.8 Implement in Matlab the second-order scheme derived in the previous exercise. Show the order of convergence numerically and compare the solutions with the first-order scheme. Comment.

Runge-Kutta methods. The main drawback of the previous second-order scheme is that it requires to know derivatives of f. This is undesirable is practice. However, the derivatives of f can also be approximated by finite differences of the form (1.7). This is the basis for the *Runge-Kutta* methods.

Exercise 1.9 Show that the following schemes are second-order:

$$\begin{aligned} g^{(2)}_{\Delta t}(n\Delta t, X) &= X + \Delta t f\Big(n\Delta t + \frac{\Delta t}{2}, X + \frac{\Delta t}{2} f(n\Delta t, X)\Big) \\ g^{(2)}_{\Delta t}(n\Delta t, X) &= X + \frac{\Delta t}{2}\Big[f(n\Delta t, X) + f\Big((n+1)\Delta t, X + \Delta t f(n\Delta t, X)\Big)\Big] \\ g^{(2)}_{\Delta t}(n\Delta t, X) &= X + \frac{\Delta t}{4}\Big[f(n\Delta t, X) + 3f\Big((n+\frac{2}{3})\Delta t, X + \frac{2\Delta t}{3} f(n\Delta t, X)\Big)\Big]. \end{aligned}$$

These schemes are called *Midpoint method*, *Modified Euler method*, and *Heun's Method*, respectively.

Exercise 1.10 Implement in Matlab the Midpoint method of the preceding exercise to solve the harmonic oscillator problem (1.5). Compare with previous discretizations.

All the methods we have seen so far are *one-step* methods, in the sense that X_{n+1} only depends on X_n. Multi-step methods are generalizations where X_{n+1} depends on X_{n-k} for $k = 0, \cdots, m$ with $m \in \mathbb{N}$. Classical multi-step methods are the Adams-Bashforth and Adams-Moulton methods. The second- and fourth-order Adams Bashforth methods are given by

$$\begin{aligned} X_{n+1} &= X_n + \frac{\Delta t}{2}(3f(n\Delta t, X_n) - f((n-1)\Delta t, X_{n-1})) \\ X_{n+1} &= X_n + \frac{\Delta t}{24}\big(55f(n\Delta t, X_n) - 59f((n-1)\Delta t, X_{n-1}) \\ &\quad +37f((n-2)\Delta t, X_{n-2}) - 9f((n-2)\Delta t, X_{n-2})\big), \end{aligned} \tag{1.23}$$

respectively.

Exercise 1.11 Implement in Matlab the Adams-Bashforth methods in (1.23) to solve the harmonic oscillator problem (1.5). Compare with previous discretizations.

2 Finite Differences for Parabolic Equations

2.1 One dimensional Heat equation

Equation. Let us consider the simplest example of a parabolic equation

$$\begin{aligned} &\frac{\partial u}{\partial t}(t,x) - \frac{\partial^2 u}{\partial x^2}(t,x) = 0, \qquad x \in \mathbb{R},\, t \in (0,T) \\ &u(0,x) = u_0(x), \qquad x \in \mathbb{R}. \end{aligned} \tag{2.1}$$

This is the one-dimensional (in space) heat equation. Note that the equation is also linear, in the sense that the map $u_0(x) \mapsto u(x,t)$ at a given time $t > 0$ is linear. In the rest of the course, we will mostly be concerned with linear equations. This equation actually admits an exact solution, given by

$$u(t,x) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\Big(- \frac{|x-y|^2}{4t} \Big) u_0(y) dy. \tag{2.2}$$

Exercise 2.1 Show that (2.2) solves (2.1).

Discretization. We now want to discretize (2.1). This time, we have two variables to discretize, time and space. The *finite difference* method consists of (i) introducing discretization points $X_n = n\Delta x$ for $n \in \mathbb{Z}$ and $\Delta x > 0$, and $T_n = n\Delta t$ for $0 \le n \le N = T/\Delta t$ and $Dt > 0$; (ii) approximating the solution $u(t,x)$ by

$$U_j^n \approx u(T_n, X_j), \qquad 1 \le n \le N,\ j \in \mathbb{Z}. \tag{2.3}$$

Notice that in practice, j is finite. However, to simplify the presentation, we assume here that we discretize the whole line $x \in \mathbb{R}$.

The time variable is very similar to what we had for ODEs: knowing what happens at $T_n = n\Delta t$, we would like to get what happens at $T_{n+1} = (n+1)\Delta t$. We therefore introduce the notation

$$\partial_t U_j^n = \frac{U_j^{n+1} - U_j^n}{\Delta t}. \tag{2.4}$$

The operator ∂_t plays the same role for the series U_j^n as $\dfrac{\partial}{\partial t}$ does for the function $u(t,x)$.

The spatial variable is quite different from the time variable. There is no privileged direction of propagation (we do not know a priori if information comes from the left or the right) as there is for the time variable (we know that information at time t will allow us to get information at time $t + \Delta t$). This intuitively explains why we introduce two types of discrete approximations for the spatial derivative:

$$\partial_x U_j^n = \frac{U_{j+1}^n - U_j^n}{\Delta x}, \qquad \text{and} \qquad \overline{\partial}_x U_j^n = \frac{U_j^n - U_{j-1}^n}{\Delta x} \tag{2.5}$$

These are the forward and backward finite difference quotients, respectively. Now in (2.1) we have a second-order spatial derivative. Since there is a priori no reason to assume that information comes from the left or the right, we choose

$$\frac{\partial^2 u}{\partial x^2}(T_n, X_j) \approx \overline{\partial}_x \partial_x U_j^n = \partial_x \overline{\partial}_x U_j^n = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}. \tag{2.6}$$

Exercise 2.2 Check the equalities in (2.6).

Notice that the discrete differentiation in (2.6) is *centered*: it is symmetric with respect to what comes from the left and what comes from the right. The finite difference scheme then reads

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} - \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} = 0. \tag{2.7}$$

This is clearly an approximation of (2.1). This is also the simplest scheme called the *explicit Euler* scheme. Notice that this can be recast as

$$U_j^{n+1} = \lambda(U_{j+1}^n + U_{j-1}^n) + (1 - 2\lambda)U_j^n, \qquad \lambda = \frac{\Delta t}{\Delta x^2}. \tag{2.8}$$

Exercise 2.3 Check this.

The numerical procedure is therefore straightforward: knowing U_j^n for $j \in \mathbb{Z}$, we calculate U_j^{n+1} for $j \in \mathbb{Z}$ for $0 \leq n \leq N-1$. There are then several questions that come to mind: is the scheme stable (i.e. does U_j^n remains bounded for $n = N$ independently of the choices of Δt and Δx)?, does it converge (is U_j^n really a good approximation of $u(n\Delta t, j\Delta x)$ as advertised)?, and how fast does it converge (what is the error between U_j^n and $u(n\Delta t, j\Delta x)$)?

Exercise 2.4 Discretizing the interval $(-10, 10)$ with N points, use a numerical approximation of (2.1) to compute $u(t, x)$ for $t = 0.01$, $t = 0.1$, $t = 1$, and $t = 10$ assuming that the initial condition is $u_0(x) = 1$ on $(-1, 1)$ and $u_0(x) = 0$ elsewhere. Implement the algorithm in Matlab.

Stability. Our first concern will be to ensure that the scheme is stable. The norm we choose to measure stability is the supremum norm. For series $U = \{U_j\}_{j\in\mathbb{Z}}$ defined on the grid $j \in \mathbb{Z}$ by U_j, we define

$$\|U\|_\infty = \sup_{j\in\mathbb{Z}} |U_j|. \tag{2.9}$$

Stability means that there exists a constant C independent of Δx, Δt and $1 \leq n \leq N = T/\Delta t$ such that

$$\|U^n\|_\infty = \sup_{j\in\mathbb{Z}} |U_j^n| \leq C\|U^0\|_\infty = C \sup_{j\in\mathbb{Z}} |U_j^0|. \tag{2.10}$$

Since $u(t, x)$ remains bounded for all times (see (2.2)), it is clear that if the scheme is unstable, it has no chance to converge. We actually have to consider two cases according as $\lambda \leq 1/2$ or $\lambda > 1/2$. When $\lambda > 1/2$, the scheme will be *unstable*! To show this we choose some initial conditions that oscillate very rapidly:

$$U_j^0 = (-1)^j \varepsilon, \qquad \varepsilon > 0.$$

Here ε is a constant that can be chosen as small as one wants. We easily verify that $\|U_j^0\|_\infty = \varepsilon$. Now straightforward calculations show that

$$U_j^1 = \big[\lambda\big((-1)^{j+1} + (-1)^{j-1}\big) + (1 - 2\lambda)(-1)^j\big]\varepsilon = (1 - 4\lambda)(-1)^j\varepsilon.$$

By induction this implies that

$$U_j^n = (1-4\lambda)^n(-1)^j\varepsilon. \tag{2.11}$$

Now assume that $\lambda > 1/2$. This implies that $\rho = -(1-4\lambda) > 1$, whence

$$\|U^n\|_\infty = \rho^n\varepsilon \to \infty \quad \text{as } n \to \infty.$$

For instance, for $n = N = T/\Delta t$, we have that $\|U^N\|_\infty$, which was supposed to be an approximation of $\sup_{x\in\mathbb{R}} |u(T,x)|$, tends to infinity as $\Delta t \to 0$. Obviously (2.10) is violated: there is no constant C independent of Δt such that (2.10) holds. In practice, what this means is that any small fluctuation in the initial data (for instance caused by computer round-off) will be amplified by the scheme exponentially fast as in (2.11) and will eventually be bigger than any meaningful information.

This behavior has to be contrasted with the case $\lambda \le 1/2$. There, stability is obvious. Since both λ and $(1-2\lambda)$ are positive, we deduce from (2.8) that

$$|U_j^{n+1}| \le \sup\{|U_{j-1}^n|, |U_j^n|, |U_{j+1}^n|\}. \tag{2.12}$$

Exercise 2.5 Check it.

So obviously,

$$\|U^{n+1}\|_\infty \le \|U^n\|_\infty \le \|U^0\|_\infty.$$

So (2.10) is clearly satisfied with $C = 1$. This property is referred to as a *maximum principle*: the discrete solution at later times ($n \ge 1$) is bounded by the maximum of the initial solution. Notice that the continuous solution of (2.1) also satisfies this property.

Exercise 2.6 *Check this using (2.2). Recall that* $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$.

Maximum principles are very important in the analysis of partial differential equations. Although we are not going to use them further here, let us still mention that it is important in general to enforce that the numerical schemes satisfy as many properties of the continuous equation as possible.

We have seen that $\lambda \le 1/2$ is important to ensure stability. In terms of Δt and Δx, this reads

$$\Delta t \le \frac{1}{2}\Delta x^2. \tag{2.13}$$

This is the simplest example of the famous *CFL* condition (after Courant, Friedrich, and Lewy who formulated it first). We shall see that explicit schemes are always *conditionally stable* for parabolic equations, whereas *implicit* schemes will be *unconditionally stable*, whence their practical interest. Notice that Δt must be quite small (of order Δx^2) in order to obtain stability.

Convergence. Now that we have stability for the Euler explicit scheme when $\lambda \le 1/2$, we have to show that the scheme converges. This is done by assuming that we have a *local* consistency of the discrete scheme with the equation, and that the solution of the equation is sufficiently *regular*. We therefore use the three same main ingredients as for ODEs: stability, consistency, regularity of the exact solution.

More specifically, let us introduce the error series $z^n = \{z_j^n\}_{j\in\mathbb{Z}}$ defined by

$$z_j^n = U_j^n - u_j^n, \qquad \text{where } u_j^n = u(n\Delta t, j\Delta x).$$

Convergence means that $\|z^n\|_\infty$ converges to 0 as $\Delta t \to 0$ and $\Delta x \to 0$ (with the constraint that $\lambda \leq 1/2$) for all $1 \leq n \leq N = T/\Delta t$. The order of convergence is how fast it converges. To simplify, we assume that $\lambda \leq 1/2$ is fixed and send $\Delta t \to 0$ (in which case $\Delta x = \sqrt{\lambda^{-1}\Delta t}$ also converges to 0). Our main ingredient is now the Taylor expansion (1.6). The consistency of the scheme consists of showing that the real dynamics are well approximated. More precisely, assuming that the solution at time $T_n = n\Delta t$ is given by u_j^n on the grid, and that $U_j^n = u_j^n$, we have to show that the real solution at time u_j^{n+1} is well approximated by U_j^{n+1}. To answer this, we calculate

$$\tau_j^n = \partial_t u_j^n - \overline{\partial}_x \partial_x u_j^n. \tag{2.14}$$

The expression τ_j^n is the truncation or local discretization error. Using Taylor expansions, we find that

$$\begin{aligned} \frac{u(t+\Delta t) - u(t)}{\Delta t} &= \frac{\partial u}{\partial t}(t) + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(t+s) \\ \frac{u(x+\Delta x) - 2u(x) + u(x-\Delta x)}{\Delta x^2} &= \frac{\partial^2 u}{\partial x^2}(x) + \frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x+h), \end{aligned}$$

for some $0 \leq s \leq \Delta t$ and $-\Delta x \leq h \leq \Delta x$.

Exercise 2.7 Check this.

Since (2.1) is satisfied by $u(t,x)$ we deduce that

$$\begin{aligned} \tau_j^n &= \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(n\Delta t + s) - \frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(j\Delta x + h) \\ &= \Delta x^2 \Big[\frac{\lambda}{2}\frac{\partial^2 u}{\partial t^2}(n\Delta t + s) - \frac{1}{12}\frac{\partial^4 u}{\partial x^4}(j\Delta x + h)\Big]. \end{aligned}$$

The truncation error is therefore of order $m = 2$ since it is proportional to Δx^2. Of course, this accuracy holds if we can bound the term in the square brackets independently of the discretization parameters Δt and Δx. Notice however that this term only involves the exact solution $u(t,x)$. All we need is that the exact solution is sufficiently regular. Again, this is independent of the discretization scheme we choose.

Let us assume that the solution is sufficiently regular (what we need is that it is of class $C^{2,4}$, i.e. that it is continuously differentiable 2 times in time and 4 times in space). We then deduce that

$$\|\tau^n\|_\infty \leq C\Delta x^2 \sim C\Delta t \tag{2.15}$$

where C is a constant independent of Δx (and Δt since λ is fixed). Since U^n satisfies the discretized equation, we deduce that

$$\partial_t z_j^n - \overline{\partial}_x \partial_x z_j^n = -\tau_j^n. \tag{2.16}$$

or equivalently that

$$z_j^{n+1} = \lambda(z_{j+1}^n + z_{j-1}^n) + (1-2\lambda)z_j^n - \Delta t \tau_j^n. \tag{2.17}$$

The above equation allows us to analyze local consistency. Indeed, let us assume that the exact solution is known on the spatial grid at time step n. This implies that $z^n = 0$, since the latter is by definition the difference between the exact solution and the approximate solution on the spatial grid. Equation (2.17) thus implies that

$$z_j^{n+1} = -\Delta t \tau_j^n,$$

so that

$$\|z^{n+1}\|_\infty \leq C\Delta t^2 \sim C\Delta t \Delta x^2, \tag{2.18}$$

thanks to (2.15). Comparing with the section on ODEs, this implies that the scheme is of order $m = 1$ in time since $\Delta t^2 = \Delta t^{m+1}$ for $m = 1$. Being of order 1 in time implies that the scheme is of order 2 is space since $\Delta t = \lambda \Delta x^2$. This will be confirmed in Theorem 2.1 below.

Let us now conclude the analysis of the error estimate and come back to (2.17). Using the stability estimate (2.12) for the homogeneous problem (because $\lambda \leq 1/2$!) and the bound (2.15), we deduce that

$$\|z^{n+1}\|_\infty \leq \|z^n\|_\infty + C\Delta t \Delta x^2.$$

This in turn implies that

$$\|z^n\|_\infty \leq nC\Delta t \Delta x^2 \leq CT\Delta x^2.$$

We summarize what we have obtain as

Theorem 2.1 *Let us assume that $u(t,x)$ is of class $C^{2,4}$ and that $\lambda \leq 1/2$. Then we have that*

$$\sup_{j\in\mathbb{Z}; 0\leq n\leq N} |u(n\Delta t, j\Delta x) - U_j^n| \leq C\Delta x^2, \tag{2.19}$$

where C is a constant independent of Δx and $\lambda \leq 1/2$.

The explicit Euler scheme is of order 2 in space. But this corresponds to a scheme of order 1 in time since $\Delta t = \lambda \Delta x^2 \leq \Delta x^2/2$. Therefore it does not converge very fast.

Exercise 2.8 Implement the explicit Euler scheme in Matlab by discretizing the parabolic equation on $x \in (-10, 10)$ and assuming that $u(t,-10) = u(t,10) = 0$. You may choose as an initial condition $u_0(x) = 1$ on $(-1,1)$ and $u_0(x) = 0$ elsewhere. Choose the discretization parameters Δt and Δx such that $\lambda = .48$, $\lambda = .5$, and $\lambda = .52$. Compare the stable solutions you obtain with the solution given by (2.2).

Exercise 2.9 Following the techniques described in Chapter 1, derive a scheme of order 4 in space (still with $\Delta t = \lambda \Delta x^2$ and λ sufficiently small that the scheme is stable). Implement this new scheme in Matlab and compare it with the explicit Euler scheme on the numerical simulation described in exercise 2.8.

2.2 Fourier transforms

We now consider a fundamental tool in the analysis of partial differential equations with homogeneous coefficients and their discretization by finite differences: Fourier transforms.

There are many ways of defining the Fourier transform. We shall use the following definition

$$[\mathcal{F}_{x\to\xi} f](\xi) \equiv \hat{f}(\xi) = \int_{-\infty}^{\infty} e^{-ix\xi} f(x) dx. \tag{2.20}$$

The inverse Fourier transform is defined by

$$[\mathcal{F}^{-1}_{\xi\to x} \hat{f}](x) \equiv f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ix\xi} \hat{f}(\xi) d\xi. \tag{2.21}$$

In the above definition, the operator $\mathcal{F}$ is applied to the function $f(x)$, which it maps to the function $\hat{f}(\xi)$. The notation can become quite useful when there are several variables and one wants to emphasize with respect to with variable Fourier transform is taken. We call the spaces of x's the physical domain and the space of ξ's the wavenumber domain, or Fourier domain. Wavenumbers are to positions what frequencies are to times. If you are more familiar with Fourier transforms of signals in time, you may want to think of "frequencies" each time you see "wavenumbers". A crucial property, which explains the terminology of "forward" and "inverse" Fourier transforms, is that

$$\mathcal{F}^{-1}_{\xi\to x} \mathcal{F}_{x\to\xi} f = f, \tag{2.22}$$

which can be recast as

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ix\xi} \int_{-\infty}^{\infty} e^{-iy\xi} f(y) dy d\xi = \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} \frac{e^{i(x-y)\xi}}{2\pi} d\xi \Big) f(y) dy.$$

This implies that

$$\int_{-\infty}^{\infty} \frac{e^{iz\xi}}{2\pi} d\xi = \delta(z),$$

the Dirac delta function since $f(x) = \int_{-\infty}^{\infty} \delta(x-y) f(y) dy$ by definition. Another way of understanding this is to realize that

$$[\mathcal{F}\delta](\xi) = 1, \qquad [\mathcal{F}^{-1} 1](x) = \delta(x). \tag{2.23}$$

The first equality can also trivially been deduced from (2.20).

Another interesting calculation for us is that

$$[\mathcal{F} \exp(-\frac{\alpha}{2} x^2)](\xi) = \sqrt{\frac{2\pi}{\alpha}} \exp(-\frac{1}{2\alpha} \xi^2). \tag{2.24}$$

In other words, the Fourier transform of a Gaussian is also a Gaussian. Notice this important fact: a very narrow Gaussian (α very large) is transformed into a very wide one (α^{-1} is very small) are vice-versa. This is related to Heisenberg's uncertainty principle in quantum mechanics, and it says that you cannot be localized in the Fourier domain when you are localized in the spatial domain and vice-versa.

Exercise 2.10 Check (2.24) using integration in the complex plane (which essentially says that $\int_{-\infty}^{\infty} e^{-\beta(x-i\rho)^2}dx = \int_{-\infty}^{\infty} e^{-\beta x^2}dx$ for $\rho \in \mathbb{R}$ by contour change since $z \to e^{-\beta z^2}$ has no pole in $\mathbb{C}$).

The simplest property of the Fourier transform is its *linearity*:

$$[\mathcal{F}(\alpha f + \beta g)](\xi) = \alpha[\mathcal{F}f](\xi) + \beta[\mathcal{F}g](\xi). \tag{2.25}$$

Other very important properties of the Fourier transform are as follows:

$$[\mathcal{F}f(x+y)](\xi) = e^{iy\xi}\hat{f}(\xi), \qquad [\mathcal{F}f(\nu x)](\xi) = \frac{1}{|\nu|}\hat{f}\Big(\frac{\xi}{\nu}\Big), \tag{2.26}$$

for all $y \in \mathbb{R}$ and $\nu \in \mathbb{R}\backslash\{0\}$.

Exercise 2.11 Check these formulas.

The first equality shows how the Fourier transform acts on translations by a factor y, the second how it acts on dilation. The former is extremely important in the analysis of PDEs and their finite difference discretization:

> *Fourier transforms replace translations in the physical domain by multiplications in the Fourier domain.*
> *More precisely, translation by y is replaced by multiplication by $e^{iy\xi}$.*

The reason this is important is that you can't beat multiplications in simplicity. To further convince you of the importance of the above assertion, consider derivatives of functions. By linearity of the Fourier transform, we get that

$$\Big[\mathcal{F}\frac{f(x+h)-f(x)}{h}\Big](\xi) = \frac{[\mathcal{F}f(x+h)](\xi) - [\mathcal{F}f(x)](\xi)}{h} = \frac{e^{ih\xi}-1}{h}\hat{f}(\xi)$$

Now let us just pass to the limit $h \to 0$ in both sides. We get

$$\hat{f}'(\xi) = [\mathcal{F}f'](\xi) = i\xi\hat{f}(\xi). \tag{2.27}$$

Again,

> *Fourier transforms replace differentiation in the physical domain by multiplication in the Fourier domain.*
> *More precisely, $\dfrac{d}{dx}$ is replaced by multiplication by $i\xi$.*

This assertion can easily been directly verified from the definition (2.20). However, it is useful to see differentiation as a limit of differences of translations. Moreover, finite differences are based on translations so we see how Fourier transforms will be useful in their analysis.

One last crucial and quite related property of the Fourier transform is that it replaces convolutions by multiplications. The convolution of two functions is given by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy = \int_{-\infty}^{\infty} f(y)g(x-y)dy. \tag{2.28}$$

The Fourier transform is then given by

$$[\mathcal{F}(f*g)](\xi) = \hat{f}(\xi)\hat{g}(\xi). \tag{2.29}$$

Exercise 2.12 Check this.

This important relation can be used to prove the equally important (everything is important in this section!) Parseval equality

$$\int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 d\xi = 2\pi \int_{-\infty}^{\infty} |f(x)|^2 dx, \tag{2.30}$$

for all complex-valued function $f(x)$ such that the above integrals make sense. We show this as follows. We first recall that $\overline{\hat{f}(\xi)} = \widehat{\overline{f}}(-\xi)$, where $\overline{f}$ denotes the complex conjugate to f. We also define $g(x) = \overline{f}(-x)$, so that $\hat{g}(\xi) = \widehat{\overline{f}}(-\xi)$ using (2.26) with $\nu = -1$. We then have

$$\begin{aligned}
&\int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi)\overline{\hat{f}(\xi)} d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi)\widehat{\overline{f}}(-\xi) d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi) g(\xi) d\xi \\
&= \int_{-\infty}^{\infty} \widehat{(f * g)}(\xi) d\xi && \text{thanks to (2.28)} \\
&= 2\pi (f * g)(0) && \text{using the definition (2.21) with } x = 0 \\
&= 2\pi \int_{-\infty}^{\infty} f(y) g(-y) dy && \text{by definition of the convolution} \\
&= 2\pi \int_{-\infty}^{\infty} f(y)\overline{f}(y) dy = 2\pi \int_{-\infty}^{\infty} |f(y)|^2 dy.
\end{aligned}$$

The Parseval equality (2.30) essentially says that there is as much energy (up to a factor 2π) in the physical domain (left-hand side in (2.30)) as in the Fourier domain (right-hand side).

Application to the heat equation. Before considering finite differences, here is an important exercise that shows how powerful Fourier transforms are to solve PDEs with constant coefficients.

Exercise 2.13 Consider the heat equation (2.1). Using the Fourier transform, show that

$$\frac{\partial \hat{u}(t,\xi)}{\partial t} = -\xi^2 \hat{u}(t,\xi).$$

Here we mean $\hat{u}(t,\xi) = [\mathcal{F}_{x\to\xi} u(t,x)](t,\xi)$. Solve this ODE. Then using (2.24) and (2.28), recover (2.2).

Application to Finite Differences. We now use the Fourier transform to analyze finite difference discretizations. To simplify we still assume that we want to solve the heat equation on the whole line $x \in \mathbb{R}$.

Let $u_0(x)$ be the initial condition for the heat equation. If we use the notation

$$U^n(j\Delta x) = U_j^n \approx u(n\Delta t, j\Delta x),$$

the explicit Euler scheme (2.7) can be recast as

$$\frac{U^{n+1}(x) - U^n(x)}{\Delta t} = \frac{U^n(x+\Delta x) + U^n(x-\Delta x) - 2U^n(x)}{\Delta x^2}, \tag{2.31}$$

where $x = j\Delta x$. As before, this scheme says that the approximation U at time $T_{n+1} = (n+1)\Delta t$ and position $X_j = j\Delta x$ depends on the values of U at time T_n and positions X_{j-1}, X_j, and X_{j+1}. It is therefore straightforward to realize that $U^n(X_j)$ depends on u_0 at all the points X_k for $j - n \leq k \leq j + n$. It is however not easy at all to get this dependence explicitly from (2.31).

The key to analyzing this dependence is to go to the Fourier domain. There, translations are replaced by multiplications, which is much simpler to deal with. Here is how we do it. We first define $U^0(x)$ for all points $x \in \mathbb{R}$, not only those points $x = X_j = j\Delta x$. One way to do so is to assume that

$$u_0(x) = U_j^0 \qquad \text{on } \Big((j - \frac{1}{2})\Delta x, (j + \frac{1}{2})\Delta x\Big]. \tag{2.32}$$

We then define $U^n(x)$ using (2.31) initialized with $U^0(x) = u_0(x)$. Clearly, we have

$$U^{n+1}(x) = (1 - 2\lambda)U^n(x) + \lambda(U^n(x + \Delta x) + U^n(x - \Delta x)). \tag{2.33}$$

We thus define $U^n(x)$ for all $n \geq 0$ and all $x \in \mathbb{R}$ by induction. If we choose as initial condition (2.32), we then easily observe that

$$U^n(x) = U_j^n \qquad \text{on } \Big((j - \frac{1}{2})\Delta x, (j + \frac{1}{2})\Delta x\Big], \tag{2.34}$$

where U_j^n is given by (2.7). So analyzing $U^n(x)$ is in some sense sufficient to get information on U_j^n.

Once $U^n(x)$ is defined for all n, we can certainly introduce the Fourier transform $\hat{U}^n(\xi)$ defined according to (2.20). However, using the translation property of the Fourier transform (2.26), we obtain that

$$\hat{U}^{n+1}(\xi) = \Big[1 + \lambda(e^{i\xi\Delta x} + e^{-i\xi\Delta x} - 2)\Big]\hat{U}^n(\xi) \tag{2.35}$$

In the Fourier domain, marching in time (going from step n to step $n + 1$) is much simpler than in the physical domain: each wavenumber ξ is multiplied by a coefficient

$$R_\lambda(\xi) = 1 + \lambda(e^{i\xi\Delta x} + e^{-i\xi\Delta x} - 2) = 1 - 2\lambda(1 - \cos(\xi\Delta x)). \tag{2.36}$$

We then have that

$$\hat{U}^n(\xi) = (R_\lambda(\xi))^n \hat{u}_0(\xi). \tag{2.37}$$

The question of the *stability* of a discretization now becomes the following: are all wavenumbers ξ stable, or equivalently is $(R_\lambda(\xi))^n$ bounded for all $1 \leq n \leq N = T/\Delta t$?

It is easy to observe that when $\lambda \leq 1/2$, $|R_\lambda(\xi)| \leq 1$ for all ξ. We then clearly have that all wavenumbers are stable since $|\hat{U}^n(\xi)| \leq |\hat{u}_0(\xi)|$. However, when $\lambda > 1/2$, we can choose values of ξ such that $\cos(\xi\Delta x)$ is close to -1. For such wavenumbers, the value of $R_\lambda(\xi)$ is less than -1 so that $\hat{U}^n$ changes sign at every new time step n with growing amplitude (see numerical simulations). Such wavenumbers are *unstable*.

2.3 Stability and convergence using Fourier transforms

In this section, we apply the theory of Fourier transform to the analysis of the finite difference of parabolic equations with constant coefficients.

As we already mentioned, an important property of Fourier transforms is that

$$\frac{\partial}{\partial x} \to i\xi,$$

that is, differentiation is replaced by multiplication by $i\xi$ in the Fourier domain. We use this to define equations in the physical domain as follows

$$\begin{aligned} &\frac{\partial u}{\partial t}(t,x) + P(D)u(t,x) = 0, \qquad t > 0,\, x \in \mathbb{R} \\ &u(0,x) = u_0(x), \qquad x \in \mathbb{R}. \end{aligned} \tag{2.38}$$

The operator $P(D)$ is a linear operator defined as

$$P(D)f(x) = \mathcal{F}^{-1}_{\xi \to x}\Big(P(i\xi)\mathcal{F}_{x\to\xi}f(x)\Big)(\xi), \tag{2.39}$$

where $P(i\xi)$ is a *function* of $i\xi$. Here D stands for $\frac{\partial}{\partial x}$. What the operator $P(D)$ does is: (i) take the Fourier transform of $f(x)$, (ii) multiply $\hat{f}(\xi)$ by the function $P(i\xi)$, (iii) take the inverse Fourier transform of the product.

The function $P(i\xi)$ is called the *symbol* of the operator $P(D)$. Such operators are called *pseudo-differential* operators.

Why do we introduce all this? The reason, as it was mentioned in the previous section, is that the analysis of differential operators simplifies in the Fourier domain. The above definition gives us a mathematical framework to build on this idea. Indeed, let us consider (2.38) in the Fourier domain. Upon taking the Fourier transform $\mathcal{F}_{x\to\xi}$ term by term in (2.38), we get that

$$\begin{aligned} &\frac{\partial \hat{u}}{\partial t}(t,\xi) + P(i\xi)\hat{u}(t,\xi) = 0, \qquad t > 0,\, \xi \in \mathbb{R}, \\ &\hat{u}(0,\xi) = \hat{u}_0(\xi), \qquad \xi \in \mathbb{R}. \end{aligned} \tag{2.40}$$

As promised, we have replaced the "complicated" pseudo-differential operator $P(D)$ by a mere multiplication by $P(i\xi)$ in the Fourier domain. For every $\xi \in \mathbb{R}$, (2.40) consists of a simple linear first-order ordinary differential equation. Notice the parallel between (2.38) and (2.40): by going to the Fourier domain, we have replaced the differentiation operator $D = \frac{\partial}{\partial x}$ by $i\xi$. The solution of (2.40) is then obviously given by

$$\hat{u}(t,\xi) = e^{-tP(i\xi)}\hat{u}_0(\xi). \tag{2.41}$$

The solution of (2.38) is then given by

$$u(t,x) = \mathcal{F}^{-1}_{\xi\to x}(e^{-tP(i\xi)}) * u_0(x). \tag{2.42}$$

Exercise 2.14 (i) Show that the parabolic equation (2.1) takes the form (2.38) with symbol $P(i\xi) = \xi^2$.
(ii) Find the equation of the form (2.38) with symbol $P(i\xi) = \xi^4$.
(iii) Show that any arbitrary partial differential equation of order $m \geq 1$ in x with constant coefficients can be written in the form (2.38) with the symbol $P(i\xi)$ a polynomial of order m.

Parabolic Symbol. In this section, we only consider real-valued symbols such that for some $M > 0$,

$$P(i\xi) \geq C|\xi|^M, \qquad \text{for all } \xi \in \mathbb{R}. \tag{2.43}$$

With a somewhat loose terminology, we refer to equations (2.38) with real-valued symbol $P(i\xi)$ satisfying the above constraint as parabolic equations. The heat equation clearly satisfies the above requirement with $M = 2$. The real-valuedness of the symbol is not necessary but we shall impose it to simplify.

What about numerical schemes? Our definition so far provides a great tool to analyze partial differential equations. It turns out it is also very well adapted to the analysis of finite difference approximations to these equations. The reason is again that finite difference operators are transformed into multiplications in the Fourier domain.

A general single-step finite difference scheme is given by

$$B_\Delta U^{n+1}(x) = A_\Delta U^n(x), \qquad n \geq 0, \tag{2.44}$$

with $U^0(x) = u_0(x)$ a given initial condition. Here, the finite difference operators are defined by

$$A_\Delta U(x) = \sum_{\alpha \in I_\alpha} a_\alpha U(x - \alpha\Delta x), \qquad B_\Delta U(x) = \sum_{\beta \in I_\beta} b_\beta U(x - \beta\Delta x), \tag{2.45}$$

where $I_\alpha \subset \mathbb{Z}$ and $I_\beta \subset \mathbb{Z}$ are finite sets of integers. Although this definition may look a little complicated at first, we need such a complexity in practice. For instance, the explicit Euler scheme is given by

$$U^{n+1}(x) = (1 - 2\lambda)U^n(x) + \lambda(U^n(x - \Delta x) + U^n(x + \Delta x)).$$

This corresponds to the coefficients

$$b_0 = 1, \quad a_{-1} = \lambda,\, a_0 = 1 - 2\lambda,\, a_1 = \lambda,$$

and all the other coefficients a_α and b_β vanish. Explicit schemes are characterized by the fact that only b_0 does not vanish. We shall see that implicit schemes, where other coefficients $b_\beta \neq 0$, are important in practice.

Notice that we are slightly changing the problem of interest. Finite difference schemes are supposed to be defined on the grid $\Delta x\mathbb{Z}$ (i.e. the points $x_m = m\Delta x$ for $m \in \mathbb{Z}$), not on the whole axis $\mathbb{R}$. Let us notice however that from the definition (2.44), the values of U^{n+1} on the grid $\Delta x\mathbb{Z}$ only depend on the values of u_0 on the same grid, and that $U^{n+1}(j\Delta x) = U_j^{n+1}$, using the notation of the preceding section. In other words $U^{n+1}(j\Delta x)$ does not depend on how we extend the initial condition u_0 originally defined on the grid $\Delta x\mathbb{Z}$ to the whole line $\mathbb{R}$. This is why we replace U_j^{n+1} by $U^{n+1}(x)$ in the sequel and only consider properties of $U^{n+1}(x)$. In the end, we should see how the properties on $U^{n+1}(x)$ translate into properties on U_j^{n+1}. This is easy to do when $U^{n+1}(x)$ is chosen as a piecewise constant function and more difficult when higher order polynomials are used to construct it. It can be done but requires careful analysis of the error made by interpolating smooth functions by polynomials. We do not consider this problem in these notes.

We now come back to (2.44). From now on, this is our definition of a finite difference scheme and we want to analyze its properties. As we already mentioned, this equation is simpler in the Fourier domain. Let us thus take the Fourier transform of each term in (2.44). What we get is

$$\Big(\sum_{\beta\in I_\beta} b_\beta e^{-i\beta\Delta x\xi}\Big)\hat{U}^{n+1}(\xi) = \Big(\sum_{\alpha\in I_\alpha} a_\alpha e^{-i\alpha\Delta x\xi}\Big)\hat{U}^n(\xi). \tag{2.46}$$

Let us introduce the symbols

$$A_\Delta(\xi) = \Big(\sum_{\alpha\in I_\alpha} a_\alpha e^{-i\alpha\Delta x\xi}\Big), \qquad B_\Delta(\xi) = \Big(\sum_{\beta\in I_\beta} b_\beta e^{-i\beta\Delta x\xi}\Big). \tag{2.47}$$

In the implicit case, we assume that

$$B_\Delta(\xi) \neq 0. \tag{2.48}$$

In the explicit case, $B_\Delta(\xi) \equiv 1$ and the above relation is obvious. We then obtain that

$$\hat{U}^{n+1}(\xi) = R_\Delta(\xi)\hat{U}^n(\xi) = B_\Delta(\xi)^{-1}A_\Delta(\xi)\hat{U}^n(\xi). \tag{2.49}$$

This implies

$$\hat{U}^n(\xi) = R_\Delta^n(\xi)\hat{u}_0(\xi). \tag{2.50}$$

The above relation completely characterizes the solution $\hat{U}^n(\xi)$. To obtain the solution $U^n(x)$, we simply have to take the inverse Fourier transform of $\hat{U}^n(\xi)$.

Exercise 2.15 Verify that for the explicit Euler scheme, we have

$$R_\Delta(\xi) = 1 - \frac{2\Delta t}{(\Delta x)^2}(1-\cos(\xi\Delta x)).$$

Notice that the difference between the exact solution $\hat{u}(n\Delta t,\xi)$ at time $n\Delta t$ and its approximation $\hat{U}^n(\xi)$ is given by

$$\hat{u}(n\Delta t,\xi) - \hat{U}^n(\xi) = \big(e^{-n\Delta t P(i\xi)} - R_\Delta^n(\xi)\big)\hat{u}_0(\xi). \tag{2.51}$$

This is what we want to analyze.

To measure convergence, we first need to define a norm, which will tell us how far or how close two functions are from one another. In finite dimensional spaces (such as $\mathbb{R}^n$ that we used for ordinary differential equations), all norms are equivalent (this is a theorem). In infinite dimensional spaces, all norms are not equivalent (this is another theorem; the equivalence of all norms on a space implies that it is finite dimensional...) So the choice of the norm matters!

Here we only consider the L^2 norm, defined on $\mathbb{R}$ for sufficiently regular complex-valued functions f by

$$\|f\| = \Big(\int_{-\infty}^{\infty} |f(x)|^2 dx\Big)^{1/2}. \tag{2.52}$$

The reason why this norm is nice is that we have the Parseval equality

$$\|\hat{f}(\xi)\| = \sqrt{2\pi}\|f(x)\|. \tag{2.53}$$

This means that the L^2 norms in the physical and the Fourier domains are equal (up to the factor $\sqrt{2\pi}$). Most norms do not have such a nice interpretation in the Fourier domain. This is one of the reasons why the L^2 norm is so popular.

Our analysis of the finite difference scheme is therefore concerned with characterizing

$$\|u(n\Delta t, x) - U^n(x)\|.$$

We just saw that this was equivalent to characterizing

$$\|\varepsilon^n(\xi)\| = \|\hat{u}(n\Delta t, \xi) - \hat{U}^n(\xi)\| = \|\Big(e^{-n\Delta t P(i\xi)} - R_\Delta^n(\xi)\Big)\hat{u}_0(\xi)\|.$$

The control that we will obtain on $\varepsilon^n(\xi)$ depends on the three classical ingredients: regularity of the solution, stability of the scheme, and consistency of the approximation.

The convergence arises as two parameters, Δt and Δx, converge to 0. To simplify the analysis, we assume that the two parameters are related by

$$\lambda = \frac{\Delta t}{\Delta x^M}, \tag{2.54}$$

where M is the order of the parabolic equation ($M = 2$ for the heat equation), and where $\lambda > 0$ is fixed.

Regularity. The regularity of the exact equation is first seen in the growth of $|e^{-\Delta t P(i\xi)}|$. Since our operator as assumed to be parabolic, we have

$$|e^{-\Delta t P(i\xi)}| \leq e^{-C\Delta t|\xi|^M}. \tag{2.55}$$

Stability. We impose that there exists a constant C independent of ξ and $1 \leq n \leq N = T/\Delta t$ such that

$$|R_\Delta^n(\xi)| \leq C, \qquad \xi \in \mathbb{R}, \quad 1 \leq n \leq N. \tag{2.56}$$

The stability constraint is clear: no wavenumber ξ can grow out of control.

Exercise 2.16 Check again that the above constraint is satisfied for the explicit Euler scheme if and only if $\lambda \leq 1/2$.

What makes parabolic equations a little special is that they have better stability properties than just having bounded symbols. The exact symbol satisfies that

$$|e^{-n\Delta t P(i\xi)}| \leq e^{-Cn\Delta t|\xi|^M}, \tag{2.57}$$

for some constant C. This relation is important because we deduce that high wavenumbers (large ξ's) are heavily damped by the operator. Parabolic equations have a very efficient *smoothing* effect. Even if high wavenumbers are present in the initial solution, they quickly disappear as time increases.

Not surprisingly, the finite difference approximation partially retains this effect. We have however to be careful. The reason is that the scheme has a given scale, Δx, at which it approximates the exact operator. For length scales much larger than Δx, we

expect the scheme to be a good approximation of the exact operator. For length of order (or smaller than) Δx however, we cannot expect the scheme to be accurate: we do not have a sufficient resolution. Remember that small scales in the spatial worlds mean high wavenumbers in the Fourier world (functions that vary on small scales oscillate fast, hence are composed of large wavenumbers). Nevertheless, for wavenumbers that are smaller than Δx^{-1}, we expect the scheme to retain some properties of the exact equation. For this reason, we impose that the operator R_Δ satisfy

$$\begin{array}{lll} |R_\Delta(\xi)| \leq 1 - \delta |\Delta x \xi|^M & \text{for} & |\Delta x \xi| \leq \gamma \\ |R_\Delta(\xi)| \leq 1 & \text{for} & |\Delta x \xi| > \gamma, \end{array} \tag{2.58}$$

where γ and δ are positive constants independent of ξ and Δx. We then have

Lemma 2.2 *We have that*

$$|R_\Delta^n(\xi)| \leq e^{-Cn|\Delta x \xi|^M} \qquad \textit{for} \quad |\Delta x \xi| \leq \gamma. \tag{2.59}$$

Proof. We deduce from (2.58) that

$$|R_\Delta^n(\xi)| \leq e^{n \ln(1-\delta|\Delta x \xi|^M)} \leq e^{-n(\delta/2)|\Delta x \xi|^M}.$$

□

Consistency. Let us now turn our attention to the consistency and accuracy of the finite difference scheme. Here again, we want to make sure that the finite difference scheme is locally a good approximation of the exact equation. In the Fourier world, this means that

$$e^{-\Delta t P(i\xi)} - R_\Delta(\xi) \tag{2.60}$$

has to converge to 0 as Δx and Δt converge to 0. Indeed, assuming that we are given the exact solution $\hat{u}(n\Delta t, \xi)$ at time $n\Delta t$, the error made by applying the approximate equation instead of the exact equation between $n\Delta t$ and $(n+1)\Delta t$ is precisely given by

$$\Big(e^{-\Delta t P(i\xi)} - R_\Delta(\xi)\Big)\hat{u}(n\Delta t, \xi).$$

Again, we have to be a bit careful. We cannot expect the above quantity to be small for all values of ξ. This is again related to the fact that wavenumbers of order of or greater than Δx^{-1} are not captured by the finite difference scheme.

Exercise 2.17 Show that in the case of the heat equation and the explicit Euler scheme, the error $e^{-\Delta t P(i\xi)} - R_\Delta(\xi)$ is not small for all wavenumbers of the form $\xi = k\pi\Delta x^{-1}$ for $k = 1, 2, \ldots$.

Quantitatively, all we can expect is that the error is small when $\xi\Delta x \to 0$ and $\Delta t \to 0$. We say that the scheme is *consistent* and *accurate of order* m if

$$|e^{-\Delta t P(i\xi)} - R_\Delta(\xi)| \leq C\Delta t \Delta x^m (1 + |\xi|^{M+m}), \quad \Delta x|\xi| < \gamma, \tag{2.61}$$

where C is a constant independent of ξ, Δx, and Δt.

Proof of convergence. Let us come back to the analysis of $\varepsilon^n(\xi)$. We have that

$$\varepsilon^n(\xi) = (e^{-\Delta t P(i\xi)} - R_\Delta(\xi)) \sum_{k=0}^{n-1} e^{-(n-1-k)\Delta t P(i\xi)} R_\Delta^k(\xi) \hat{u}_0(\xi),$$

using $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^{n-1-k} b^k$. Let us first consider the case $|\xi| \leq \gamma \Delta x^{-1}$. We deduce from (2.57), (2.59), and (2.61), that

$$|\varepsilon^n(\xi)| \leq C \Delta t \Delta x^m (1 + |\xi|^{M+m}) n e^{-Cn\Delta t |\xi|^M} |\hat{u}_0(\xi)|. \tag{2.62}$$

Now remark that

$$n\Delta t |\xi|^M e^{-Cn\Delta t|\xi|^M} \leq C'.$$

Since $n\Delta t \leq T$, this implies that

$$|\varepsilon^n(\xi)| \leq C \Delta x^m (1 + |\xi|^m) |\hat{u}_0(\xi)|. \tag{2.63}$$

So far, the above inequality holds for $|\xi| \leq \gamma \Delta x^{-1}$. However, for $|\xi| \geq \gamma \Delta x^{-1}$, we have thanks to the *stability property* (2.56) that

$$|\varepsilon^n(\xi)| \leq 2|\hat{u}_0(\xi)| \leq C \Delta x^m (1 + |\xi|^m) |\hat{u}_0(\xi)|. \tag{2.64}$$

So (2.63) actually holds for every ξ. This implies that

$$\|\varepsilon^n\| = \Big(\int_{\mathbb{R}} |\varepsilon^n(\xi)|^2 d\xi \Big)^{1/2} \leq C \Delta x^m \Big(\int_{\mathbb{R}} (1 + |\xi|^{2m}) |\hat{u}_0(\xi)|^2 d\xi \Big)^{1/2}. \tag{2.65}$$

This shows that $\|\varepsilon^n\|$ is of order Δx^m provided that

$$\int_{\mathbb{R}} (1 + |\xi|^{2m}) |\hat{u}_0(\xi)|^2 d\xi \Big)^{1/2} < \infty. \tag{2.66}$$

The above inequality implies that $\hat{u}_0(\xi)$ decays sufficiently fast as $\xi \to \infty$. This is actually equivalent to imposing that $u_0(x)$ is sufficiently *regular*.

The Hilbert spaces $H^m(\mathbb{R})$. For a function $v(x)$, we denote by $v^{(n)}(x)$ its nth derivative. We introduce the sequence of functional spaces $H^m = H^m(\mathbb{R})$ of functions whose derivatives up to order m are square-integrable. The norm of H^m is defined by

$$\|v\|_{H^m} = \Big(\sum_{k=0}^{m} \int_{\mathbb{R}} |v^{(n)}(x)|^2 dx \Big)^{1/2}. \tag{2.67}$$

Notice that $H^0 = L_2$, the space of square-integrable functions. It turns out that a function $v(x)$ belongs to H^m if and only if its Fourier transform $\hat{v}(\xi)$ satisfies that

$$\Big(\int_{\mathbb{R}} (1 + |\xi|^{2m}) |\hat{v}(\xi)|^2 d\xi \Big)^{1/2} < \infty. \tag{2.68}$$

Exercise 2.18 (difficult) Prove it. The proof is based on the fact that differentiation in the physical domain is replaced by multiplication by $i\xi$ in the Fourier domain.

Notice that in the above inequality, we can choose $m \in \mathbb{R}$ and not only $m \in \mathbb{N}^*$. Thus (2.68) can be used as a definition for the space H^m with $m \in \mathbb{R}$. The space H^m is still the space of functions with m square-integrable derivatives. Only the number of derivatives can be "fractional", or even negative!

Exercise 2.19 Calculate the Fourier transform of the function v defined by $v(x) = 1$ on $(-1, 1)$ and $v(x) = 0$ elsewhere. Using (2.68), show that the function v belongs to every space H^m with $m < 1/2$. This means that functions that are piecewise smooth (with discontinuities) have a little less than half of a derivative in L^2 (you can differentiate such functions almost $1/2$ times)!

The relationship between (2.67) and (2.68) can be explained as follows:

$v(x)$ is smooth	$\Longleftrightarrow$	$\hat{v}(\xi)$ decays fast
in the physical domain		in the Fourier domain.

This is another very general and very important property of Fourier transforms. The equivalence between (2.67) and (2.68) is one manifestation of this "rule".

Main result of the section. Summarizing all of the above, we have obtained the following result.

Theorem 2.3 *Let us assume that $P(i\xi)$ and $R_\Delta(\xi)$ are parabolic, in the sense that (2.43) and (2.58) are satisfied and that $R_\Delta(\xi)$ is of order m, i.e. (2.61) holds. Assuming that the initial condition $u_0 \in H^m$, we obtain that*

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^m \|u_0\|_{H^m}, \qquad \textit{for all } 0 \leq n \leq N = T/\Delta t. \tag{2.69}$$

We recall that $\Delta t = \lambda \Delta x^M$ so that in the above theorem, the scheme is of order m in space and of order m/M in time.

Exercise 2.20 (difficult) Follow the proof of Theorem 2.3 assuming that (2.43) and (2.58) are replaced by $P(i\xi) \geq 0$ and $R_\Delta(\xi) \leq 1$. Show that (2.69) should be replaced by

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^m \|u_0\|_{H^{m+M}}, \qquad \text{for all } 0 \leq n \leq N = T/\Delta t. \tag{2.70}$$

Loosely speaking, what this means is that we need the initial condition to be more regular if the equation is not regularizing. You can compare this with Theorem 2.1: to obtain a convergence of order 2 in space, we have to assume that the solution $u(t, x)$ is of order C^4 is space. This is consistent with the above result with $M = 2$ and $m = 2$.

Exercise 2.21 (very difficult). Assuming that (2.58) is replaced by $R_\Delta(\xi) \leq 1$, show that (2.69) should be replaced by

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^m \|u_0\|_{H^{m+\varepsilon}}, \qquad \text{for all } 0 \leq n \leq N = T/\Delta t. \tag{2.71}$$

Here ε is an arbitrary positive constant. Of course, the constant C depends on ε. What this result means is that if we loose the parabolic property of the discrete scheme, we still have the same order of convergence provided that the initial condition is slightly more regular than being in H^m.

Exercise 2.22 Show that any stable scheme that is convergent of order m is also convergent of order αm for all $0 \leq \alpha \leq 1$. Deduce that (2.69) in Theorem 2.3 can then be replaced by

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^{\alpha m}\|u_0\|_{H^{\alpha m}}, \qquad \text{for all } 0 \leq n \leq N = T/\Delta t. \tag{2.72}$$

The interest of this result is the following: when the initial data u_0 is not sufficiently regular to be in H^m but sufficiently regular to be in $H^{\alpha m}$ with $0 < \alpha < 1$, we still have convergence of the finite difference scheme as $\Delta x \to 0$; however the rate of convergence is slower.

Exercise 2.23 Show that everything we have said in this section holds if we replace (2.43) and (2.58) by

$$\text{Re } P(i\xi) \geq C(|\xi|^M - 1) \qquad \text{and} \qquad |R_\Delta(\xi)| \leq 1 - \delta|\Delta x\xi|^M + C\Delta t, \quad |\Delta x\xi| \leq \gamma,$$

respectively. Here, Re stands for "real part". The above hypotheses are sufficiently general to cover most practical examples of parabolic equations.

Exercise 2.24 Show that schemes accurate of order $m > 0$ so that (2.61) holds and verifying $|R_\Delta(\xi)| \leq 1$ are stable in the sense that

$$|R_\Delta(\xi)| \leq (1 + C\Delta x^m \Delta t) - \delta|\Delta x\xi|^M, \quad |\Delta x\xi| \leq \gamma.$$

The last exercises show that schemes that are stable in the "classical" sense, i.e., that verify $|R_\Delta(\xi)| < 1$, and are consistent with the smoothing parabolic symbol, are themselves smoothing.

2.4 Application to the θ schemes

We are now ready to analyze the family of schemes for the heat equation (i.e. $P(i\xi) = \xi^2$) called the θ schemes.

The θ scheme is defined by

$$\partial_t U^n(x) = \theta\partial_x\overline{\partial}_x U^{n+1}(x) + (1-\theta)\partial_x\overline{\partial}_x U^n(x). \tag{2.73}$$

Recall that the finite difference operators are defined by (2.4) and (2.5). When $\theta = 0$, we recover the explicit Euler scheme. For $\theta = 1$, the scheme is called the fully implicit Euler scheme. For $\theta = 1/2$, the scheme is called the Crank-Nicolson scheme. We recall that $\Delta t = \lambda\Delta x^2$.

Exercise 2.25 Show that the symbol of the θ scheme is given by

$$R_\Delta(\xi) = \frac{1 - 2(1-\theta)\lambda(1 - \cos(\Delta x\xi))}{1 + 2\theta\lambda(1 - \cos(\Delta x\xi))}. \tag{2.74}$$

Let us first consider the stability of the θ method. Let us assume that $0 \leq \theta \leq 1$. We observe that $R_\Delta(\xi) \leq 1$. The stability requirement is therefore that

$$\min_\xi R_\Delta(\xi) \geq -1.$$

Exercise 2.26 Show that the above inequality holds if and only if

$$(1-2\theta)\lambda \leq \frac{1}{2}.$$

This shows that the method is *unconditionally* L^2-stable when $\theta \geq 1/2$, i.e. that the method is stable independently of the choice of λ. When $\theta < 1/2$, stability only holds if and only if

$$\lambda \leq \frac{1}{2(1-2\theta)}.$$

Let us now consider the accuracy of the θ method. Obviously we need to show that (2.61) holds for some $m > 0$.

Exercise 2.27 Show that

$$e^{-\Delta t\xi^2} = 1 - \Delta t\xi^2 + \frac{1}{2}\Delta t^2\xi^4 - \frac{1}{6}\Delta t^3\xi^6 + O(\Delta t^4\xi^8), \tag{2.75}$$

and that

$$R_\Delta(\xi) = 1-\Delta t\xi^2+\Delta t\Delta x^2(\frac{1}{12}+\lambda\theta)\xi^4-\frac{\Delta t^2\Delta x^4}{360}(1+60\lambda\theta+360\lambda^2\theta^2)+O(\Delta t^4\xi^8), \tag{2.76}$$

as $\Delta t \to 0$ and $\Delta x\xi \to 0$.

Exercise 2.28 Show that the θ scheme is always at least second-order in space (first-order in time).
Show that the scheme is of order 4 in space if and only if

$$\theta = \frac{1}{2} - \frac{1}{12\lambda}.$$

Show that the scheme is of order 6 in space if in addition,

$$1 + 60\lambda\theta + 360\lambda^2\theta^2 = 60\lambda^2, \qquad \text{i.e., } \lambda = \frac{1}{10}\sqrt{5}.$$

Implementation of the θ scheme. The above analysis shows that the θ scheme is unconditionally stable when $\theta \geq 1/2$. This very nice stability property comes however at a price: the calculation of U^{n+1} from U^n is no longer explicit.

Exercise 2.29 Show that the θ scheme can be put in the form (2.44) with

$$b_{-1} = b_1 = -\theta\lambda, \quad b_0 = 1 + 2\lambda\theta, \qquad a_{-1} = a_1 = (1-\theta)\lambda, \quad a_0 = 1 - 2(1-\theta)\lambda$$

Since the coefficient b_1 and b_{-1} do not vanish, the scheme is no longer explicit. Let us come back to the solution U_j^n defined on the grid $\Delta x\mathbb{Z}$. The operators A_Δ and B_Δ in (2.45) are now infinite matrices such that

$$B_\Delta U^{n+1} = A_\Delta U^n$$

where the infinite vector $U^n = (U_j^n)_{j\in\mathbb{Z}}$. In matrix form, we obtain that the operator B_Δ for the θ scheme is given by the infinite matrix

$$B_\Delta = \begin{pmatrix} 1+2\theta\lambda & -\theta\lambda & 0 & 0 & 0 \\ -\theta\lambda & 1+2\theta\lambda & -\theta\lambda & 0 & 0 \\ 0 & -\theta\lambda & 1+2\theta\lambda & -\theta\lambda & 0 \\ 0 & 0 & -\theta\lambda & 1+2\theta\lambda & -\theta\lambda \\ 0 & 0 & 0 & -\theta\lambda & 1+2\theta\lambda \end{pmatrix}. \tag{2.77}$$

Similarly, the matrix A_Δ is given by

$$A_\Delta = \begin{pmatrix} 1-2(1-\theta)\lambda & (1-\theta)\lambda & 0 & 0 & 0 \\ (1-\theta)\lambda & 1-2(1-\theta)\lambda & (1-\theta)\lambda & 0 & 0 \\ 0 & (1-\theta)\lambda & 1-2(1-\theta)\lambda & (1-\theta)\lambda & 0 \\ 0 & 0 & (1-\theta)\lambda & 1-2(1-\theta)\lambda & (1-\theta)\lambda \\ 0 & 0 & 0 & (1-\theta)\lambda & 1-2(1-\theta)\lambda \end{pmatrix}. \tag{2.78}$$

The vector U^{n+1} is thus given by

$$U^{n+1} = (B_\Delta)^{-1} A_\Delta U^n. \tag{2.79}$$

This implies that in any numerical implementation of the θ scheme where $\theta > 0$, we need to invert the system $B_\Delta U^{n+1} = A_\Delta U^n$ (which is usually much faster than constructing the matrix $(B_\Delta)^{-1}$ directly). In practice, we cannot use infinite matrices obviously. However we can use the $N \times N$ matrices above on an interval of size $N\Delta x$ given. The above matrices correspond to imposing that the solution vanishes at the boundary of the interval (Dirichlet boundary conditions).

Exercise 2.30 Implement the θ scheme in Matlab for all possible values of $0 \leq \theta \leq 1$ and $\lambda > 0$.
Verify on a few examples with different values of $\theta < 1/2$ that the scheme is unstable when λ is too large.
Verify numerically (by dividing an initial spatial mesh by a factor 2, then 4) that the θ scheme is of order 2, 4, and 6 in space for well chosen values of θ and λ. To obtain these orders of convergence, make sure that the initial condition is sufficiently smooth. For non-smooth initial conditions, show that the order of convergence decreases and compare with the theoretical predictions of Theorem 2.3.

3 Finite Differences for Elliptic Equations

The prototype of elliptic equation we consider is the following two-dimensional diffusion equation with absorption coefficient

$$-\Delta u(\mathbf{x}) + \sigma u(\mathbf{x}) = f(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^2. \tag{3.1}$$

Here, $\mathbf{x} = (x, y) \in \mathbb{R}^2$, $f(\mathbf{x})$ is a given source term, σ is a given positive absorption coefficient, and

$$\Delta = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is the Laplace operator. We assume that our domain has no boundaries to simplify. The above partial differential equation can be analyzed by Fourier transform. In two-space dimensions, the Fourier transform is defined by

$$[\mathcal{F}_{\mathbf{x}\to\boldsymbol{\xi}} f](\boldsymbol{\xi}) \equiv \hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^2} e^{-i\mathbf{x}\cdot\boldsymbol{\xi}} f(\mathbf{x}) d\mathbf{x}. \tag{3.2}$$

The inverse Fourier transform is then defined by

$$[\mathcal{F}^{-1}_{\boldsymbol{\xi}\to\mathbf{x}} \hat{f}](\mathbf{x}) \equiv f(\mathbf{x}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{i\mathbf{x}\cdot\boldsymbol{\xi}} \hat{f}(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{3.3}$$

We still verify that the Fourier transform replaces translations and differentiation by multiplications. More precisely, we have

$$[\mathcal{F} f(\mathbf{x} + \mathbf{y})](\boldsymbol{\xi}) = e^{i\mathbf{y}\cdot\boldsymbol{\xi}} \hat{f}(\boldsymbol{\xi}), \qquad [\mathcal{F}\nabla f(\mathbf{x})](\boldsymbol{\xi}) = i\boldsymbol{\xi}\hat{f}(\boldsymbol{\xi}). \tag{3.4}$$

Notice here that both ∇ and $\boldsymbol{\xi}$ are 2-vectors. This implies that the operator $-\Delta$ is replaced by a multiplication by $|\boldsymbol{\xi}|^2$ (Check it!). Upon taking the Fourier transform in both terms of the above equation, we obtain that

$$(\sigma + |\boldsymbol{\xi}|^2)\hat{u}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}), \qquad \boldsymbol{\xi} \in \mathbb{R}^2. \tag{3.5}$$

The solution to the diffusion equation is then given in the Fourier domain by

$$\hat{u}(\boldsymbol{\xi}) = \frac{\hat{f}(\boldsymbol{\xi})}{\sigma + |\boldsymbol{\xi}|^2}. \tag{3.6}$$

It then remains to apply the inverse Fourier transform to the above equality to obtain $u(\mathbf{x})$. Since the inverse Fourier transform of a product is the convolution of the inverse Fourier transforms (as in 1D), we deduce that

$$\begin{aligned} u(\mathbf{x}) &= \Big(\mathcal{F}^{-1}\frac{1}{\sigma + |\boldsymbol{\xi}|^2}\Big)(\mathbf{x}) * f(\mathbf{x}) \\ &= CK_0(\sqrt{\sigma}|\mathbf{x}|) * f(\mathbf{x}), \end{aligned}$$

where K_0 is the modified Bessel function of order 0 of the second kind and C is a normalization constant.

This shows again how powerful Fourier analysis is to solve partial differential equations with constant coefficients. We could define mode general equations of the form

$$P(D)u(x) = f(x) \qquad x \in \mathbb{R}^2,$$

where $P(D)$ is an operator with symbol $P(i\boldsymbol{\xi})$. In the case of the diffusion equation, we would have that $P(i\boldsymbol{\xi}) = \sigma + |\boldsymbol{\xi}|^2$, i.e. $P(i\xi)$ is a parabolic symbol of order 2 using the terminology of the preceding section. All we will see in this section applies to quite general symbols $P(i\boldsymbol{\xi})$, although we shall consider only $P(i\boldsymbol{\xi}) = \sigma + |\boldsymbol{\xi}|^2$ to simplify.

Discretization. How do we solve (3.1) by finite differences? We can certainly replace the differentials by approximated differentials as we did in 1D. This is the finite difference approach. Let us consider the grid of points $\mathbf{x}_{ij} = (i\Delta x, j\Delta x)$ for $i, j \in \mathbb{Z}$ and $\Delta x > 0$ given. We then define

$$U_{ij} \approx u(\mathbf{x}_{ij}), \qquad f_{ij} = f(\mathbf{x}_{ij}). \tag{3.7}$$

Using a second-order scheme, we can then replace (3.1) by

$$-\big(\partial_x\overline{\partial}_x + \partial_y\overline{\partial}_y\big)U_{ij} + \sigma U_{ij} = f_{ij}, \qquad (i,j) \in \mathbb{Z}^2, \tag{3.8}$$

where the operator ∂_x is defined by

$$\partial_x U_{ij} = \frac{U_{i+1,j} - U_{ij}}{\Delta x}, \tag{3.9}$$

and $\overline{\partial}_x$, ∂_y, and $\overline{\partial}_y$ are defined similarly.

Assuming as in the preceding section that the finite difference operator is defined in the whole space $\mathbb{R}^2$ and not only on the grid $\{\mathbf{x}_{ij}\}$, we have more generally that

$$BU(\mathbf{x}) = f(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^2, \tag{3.10}$$

where the operator B is a sum of translations defined by

$$BU(\mathbf{x}) = \sum_{\boldsymbol{\beta} \in I_{\boldsymbol{\beta}}} b_{\boldsymbol{\beta}} U(\mathbf{x} - \Delta x\boldsymbol{\beta}). \tag{3.11}$$

Here, $I_{\boldsymbol{\beta}}$ is a finite subset of $\mathbb{Z}^2$, i.e. $\boldsymbol{\beta} = (\beta_1, \beta_2)$, where β_1 and β_2 are integer. Upon taking Fourier transform of (3.11), we obtain

$$\Big(\sum_{\boldsymbol{\beta} \in I_{\boldsymbol{\beta}}} b_\beta e^{-i\Delta x\boldsymbol{\beta}\cdot\boldsymbol{\xi}}\Big)\hat{U}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}). \tag{3.12}$$

Defining

$$R_\Delta(\boldsymbol{\xi}) = \Big(\sum_{\boldsymbol{\beta} \in I_{\boldsymbol{\beta}}} b_\beta e^{-i\Delta x\boldsymbol{\beta}\cdot\boldsymbol{\xi}}\Big), \tag{3.13}$$

we then obtain that the finite difference solution is given in the Fourier domain by

$$\hat{U}(\boldsymbol{\xi}) = \frac{1}{R_\Delta(\boldsymbol{\xi})}\hat{f}(\boldsymbol{\xi}), \tag{3.14}$$

and that the difference between the exact and the discrete solutions is given by

$$\hat{u}(\boldsymbol{\xi}) - \hat{U}(\boldsymbol{\xi}) = \Big(\frac{1}{P(i\boldsymbol{\xi})} - \frac{1}{R_\Delta(\boldsymbol{\xi})}\Big)\hat{f}(\boldsymbol{\xi}). \tag{3.15}$$

For the second-order scheme introduced above, we obtain

$$R_\Delta(\boldsymbol{\xi}) = \sigma + \frac{2}{\Delta x^2}\big(2 - \cos(\Delta x \xi_x) - \cos(\Delta x \xi_y)\big). \tag{3.16}$$

For $\Delta x|\boldsymbol{\xi}| \leq \pi$, we deduce from Taylor expansions that

$$|R_\Delta(\boldsymbol{\xi}) - P(i\boldsymbol{\xi})| \leq C\Delta x^2 |\boldsymbol{\xi}|^4.$$

For $\Delta x|\boldsymbol{\xi}| \geq \pi$, we also obtain that

$$|R_\Delta(\boldsymbol{\xi})| \leq \frac{C}{\Delta x^2} \leq C\Delta x^2 |\boldsymbol{\xi}|^4,$$

and

$$|P(i\boldsymbol{\xi})| \leq C\Delta x^2 |\boldsymbol{\xi}|^4.$$

This implies that

$$|R_\Delta(\boldsymbol{\xi}) - P(i\boldsymbol{\xi})| \leq C\Delta x^2 |\boldsymbol{\xi}|^4$$

for all $\boldsymbol{\xi} \in \mathbb{R}^2$. Since $R_\Delta(\boldsymbol{\xi}) \geq \sigma$ and $P(i\boldsymbol{\xi}) \geq C(1 + |\boldsymbol{\xi}|^2)$, we then easily deduce that

$$\Big|\frac{R_\Delta(\boldsymbol{\xi}) - P(i\boldsymbol{\xi})}{R_\Delta(\boldsymbol{\xi})P(i\boldsymbol{\xi})}\Big| \leq C\Delta x^2(1 + |\boldsymbol{\xi}|^2). \tag{3.17}$$

From this, we deduce that

$$|\hat{u}(\boldsymbol{\xi}) - \hat{U}(\boldsymbol{\xi})| \leq C\Delta x^2(1 + |\boldsymbol{\xi}|^2)|\hat{f}(\boldsymbol{\xi})|.$$

Let us now square the above inequality and integrate. Upon taking the square root of the integrals, we obtain that

$$\|\hat{u}(\boldsymbol{\xi}) - \hat{U}(\boldsymbol{\xi})\| \leq C\Delta x^2 \Big(\int_{\mathbb{R}^2} (1 + |\boldsymbol{\xi}|^2)^2 |\hat{f}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}\Big)^{1/2}. \tag{3.18}$$

The above integral is bounded provided that $f(\mathbf{x}) \in H^2(\mathbb{R}^2)$, using the definitions (2.67)-(2.68), which also hold in two space dimensions. From the Parseval equality (equation (2.53) with $\sqrt{2\pi}$ replaced by 2π in two space dimensions), we deduce that

$$\|u(\mathbf{x}) - U(\mathbf{x})\| \leq C\Delta x^2 \|f(\mathbf{x})\|_{H^2(\mathbb{R}^2)}. \tag{3.19}$$

This is the final result of this section: provided that the source term is sufficiently regular (its second-order derivatives are square-integrable), the error made by the finite difference approximation is of order Δx^2. The scheme is indeed second-order.

Exercise 3.1 Notice that the bound in (3.17) would be replaced by $C\Delta x^2(1 + |\boldsymbol{\xi}|^4)$ if only the Taylor expansion were used. The reason why (3.17) holds is because the exact symbol $P(i\boldsymbol{\xi})$ damps high frequencies (i.e. is bounded from below by $C(1 + |\boldsymbol{\xi}|^2)$. Show that if this damping is not used in the proof, the final result (3.19) should be replaced by

$$\|u(\mathbf{x}) - U(\mathbf{x})\| \leq C\Delta x^2 \|f(\mathbf{x})\|_{H^4(\mathbb{R}^2)}.$$

Exercise 3.2 (moderately difficult) Show that

$$\Big|\frac{R_\Delta(\boldsymbol{\xi}) - P(i\boldsymbol{\xi})}{R_\Delta(\boldsymbol{\xi})P(i\boldsymbol{\xi})}\Big| \leq C\Delta x^m(1 + |\boldsymbol{\xi}|^m),$$

for all $0 \leq m \leq 2$. [Hint: Show that $(R_\Delta(\boldsymbol{\xi}))^{-1}$ and $(P(i\xi))^{-1}$ are bounded and separate $|\boldsymbol{\xi}|\Delta x \geq 1$ and $|\boldsymbol{\xi}|\Delta x \leq 1$].
Deduce that

$$\|u(\mathbf{x}) - U(\mathbf{x})\| \leq C\Delta x^m \|f(\mathbf{x})\|_{H^m(\mathbb{R}^2)},$$

for $0 \leq m \leq 2$. We deduce from this result that the error $u(\mathbf{x}) - U(\mathbf{x})$ still converges to 0 when f is less regular than $H^2(\mathbb{R}^2)$. However, the convergence is slower and no longer of order Δx^2.

Exercise 3.3 Implement the Finite Difference algorithm in Matlab on a square of finite dimension (impose that the solution vanishes at the boundary of the domain). This implementation is not trivial. It requires constructing and inverting a matrix as was done in (2.79).
Solve the diffusion problem with a smooth function f and show that the order of convergence is 2. Show that the order of convergence decreases when f is less regular.

4 Finite Differences for Hyperbolic Equations

After parabolic and elliptic equations, we turn to the third class of important linear equations: hyperbolic equations. The simplest hyperbolic equation is the one-dimensional transport equation

$$\begin{aligned} &\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \qquad t > 0,\, x \in \mathbb{R} \\ &u(0,x) = u_0(x), \qquad x \in \mathbb{R}. \end{aligned} \tag{4.1}$$

Here $u_0(x)$ is the initial condition. When a is constant, the solution to the above equation is easily found to be

$$u(t,x) = u_0(x - at). \tag{4.2}$$

This means that the initial data is simply translated by a speed a. The solution can also be obtained by Fourier transforms. Indeed we have

$$\frac{\partial \hat{u}}{\partial t}(t,\xi) + ia\xi\hat{u}(t,\xi) = 0, \tag{4.3}$$

which gives

$$\hat{u}(t,\xi) = \hat{u}_0(\xi)e^{-iat\xi}.$$

Now since multiplication by a phase in the Fourier domain corresponds to a translation in the physical domain, we obtain (4.2).

Here is an important remark. Notice that (4.3) can be written as (2.40) with $P(i\xi) = ia\xi$. Notice that this symbol does not have the properties imposed in section 2. In particular, the symbol $P(i\xi)$ is *purely imaginary* (i.e. its real part vanishes). This corresponds to a completely different behavior of the PDE solutions: instead of regularizing initial discontinuities, as parabolic equations do, hyperbolic equations translate and possibly modify the shape of discontinuities, but do not regularize them. This behavior will inevitably show up when we discretize hyperbolic equations.

Let us now consider finite difference discretizations of the above transport equation. The framework introduced in section 2 still holds. The single-step finite difference scheme is given by (2.44). After Fourier transforms, it is defined by (2.49). The L^2 norm of the difference between the exact and approximate solutions is then given by

$$\|u(n\Delta t,x) - U^n(x)\| = \frac{1}{\sqrt{2\pi}}\|\varepsilon^n(\xi)\| = \frac{1}{\sqrt{2\pi}}\big\|\big(e^{-n\Delta t P(i\xi)} - R_\Delta^n(\xi)\big)\hat{u}_0(\xi)\big\|.$$

Here, we recall that $P(i\xi) = ia\xi$. The mathematical analysis of finite difference schemes for hyperbolic equations is then not very different from for parabolic equations. Only the results are very different: schemes that may be natural for parabolic equations will no longer be so for hyperbolic schemes.

Stability. One of the main steps in controlling the error term ε^n is to show that the scheme is stable. By stability, we mean here that

$$|R_\Delta^n(\xi)| = |R_\Delta(\xi)|^n \le C, \qquad n\Delta t \le T.$$

Let us consider the stability of classical schemes. The first idea that comes to mind to discretize (4.1) is to replace the time derivative by ∂_t and the spatial derivative by ∂_x. Spelling out the details, this gives

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{\Delta x}(U_{j+1}^n - U_j^n).$$

After Fourier transform (and extension of the scheme to the whole line $x \in \mathbb{R}$), we deduce that

$$R_\Delta(\xi) = 1 - \nu(e^{i\xi\Delta x} - 1) = [1 - \nu(\cos(\xi\Delta x) - 1)] - i\nu\sin(\xi\Delta x),$$

where we have defined

$$\nu = \frac{a\Delta t}{\Delta x}. \tag{4.4}$$

We deduce that

$$|R_\Delta(\xi)|^2 = 1 + 2\nu(1+\nu)(1 - \cos(\xi\Delta x)).$$

When $a > 0$, which implies that $\nu > 0$, we deduce that the scheme is always *unstable* since $|R_\Delta(\xi)|^2 > 1$ for $\cos(\xi\Delta x) \leq 0$ for instance. When $\nu < -1$, we again deduce that $|R_\Delta(\xi)|^2 > 1$ since $\nu(1+\nu) > 0$. It remains to consider $-1 \leq \nu \leq 0$. There, $\nu(1+\nu) \leq 0$ and we then deduce that $|R_\Delta(\xi)|^2 \leq 1$ for all wavenumbers ξ. This implies stability.

So the scheme is stable if and only if $-1 \leq \nu \leq 0$. This implies that $|a|\Delta t \leq \Delta x$. Again, the time step cannot be chosen too large. This is referred to as a CFL condition. Notice however, that this also implies that $a < 0$. This can be understood as follows. In the transport equation, information propagates from the right to the left when $a < 0$. Since the scheme approximates the spatial derivative by ∂_x, it is asymmetrical and uses some information to the right (at the point $X_{j+1} = X_j + \Delta x$) of the current point X_j. For $a < 0$, it gets the correct information where it comes from. When $a > 0$, the scheme still "looks" for information coming from the right, whereas physically it comes from the left.

This leads to the definition of the *upwind scheme*, defined by

$$U_j^{n+1} = \begin{cases} U_j^n - \nu(U_j^n - U_{j-1}^n) & \text{when} \quad a > 0 \\ U_j^n - \nu(U_{j+1}^n - U_j^n) & \text{when} \quad a < 0. \end{cases} \tag{4.5}$$

Exercise 4.1 Check that the upwind scheme is stable provided that $|a|\Delta t \leq \Delta x$. Implement the scheme in Matlab.

The above result shows that we have to "know" which direction the information comes from to construct our scheme. For one-dimensional problems, this is relatively easy since only the sign of a is involved. In higher dimensions, knowing where the information comes from is much more complicated. A tempting solution to this difficulty is to average over the operators ∂_x and $\overline{\partial}_x$:

$$U_j^{n+1} = U_j^n - \frac{\nu}{2}(U_{j+1}^n - U_{j-1}^n).$$

Exercise 4.2 Show that the above scheme is always *unstable*. Verify the instability in Matlab.

A more satisfactory answer to the question has been answered by Lax and Wendroff. The *Lax-Wendroff scheme* is defined by

$$U_j^{n+1} = \frac{1}{2}\nu(1+\nu)U_{j-1}^n + (1-\nu^2)U_j^n - \frac{1}{2}\nu(1-\nu)U_{j+1}^n. \tag{4.6}$$

Exercise 4.3 Check that

$$|R_\Delta(\xi)|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4\frac{\xi\Delta x}{2}.$$

Deduce that the Lax-Wendroff scheme is stable when $|\nu| \leq 1$.
Implement the Lax-Wendroff scheme in Matlab.

Consistency. We have introduced two stable schemes, the upwind scheme and the Lax-Wendroff scheme. It remains to analyze their convergence properties.

Consistency consists of analyzing the local properties of the discrete scheme and making sure that the scheme captures the main trends of the continuous equation. So assuming that $u(n\Delta t, X_j)$ is known at time $n\Delta t$, we want to see what it becomes under the discrete scheme. For the upwind scheme with $a > 0$ to simplify, we want to analyze

$$\Big(\partial_t + a\overline{\partial}_x\Big)u(T_n, X_j).$$

The closer the above quantity to 0 (which it would be if the discrete scheme were replaced by the continuous equation), the higher the order of convergence of the scheme. This error is obtained assuming that the exact solution is sufficiently *regular* and by using Taylor expansions.

Exercise 4.4 Show that

$$\Big(\partial_t + a\overline{\partial}_x\Big)u(T_n, X_j) = -\frac{1}{2}(1-\nu)a\Delta x\frac{\partial^2 u}{\partial x^2}(T_n, X_j) + O(\Delta x^2).$$

This shows that the upwind scheme is first-order when $\nu < 1$. Notice that first-order approximations involve second-order derivatives of the exact solution. This is similar to the parabolic case and justifies the following definition in the Fourier domain (remember that second-order derivative in the physical domain means multiplication by $-\xi^2$ in the Fourier domain).

We say that the scheme is convergent of order m if

$$|e^{-i\Delta t P(i\xi)} - R_\Delta(\xi)| \leq C\Delta t\Delta x^m(1+|\xi|^{m+1}), \tag{4.7}$$

for Δt and $\Delta x|\xi|$ sufficiently small (say $\Delta x|\xi| \leq \pi$). Notice that Δt and Δx are again related through (4.4).

Notice the parallel with (2.61): we have simply replaced M by 1. The difference between parabolic and hyperbolic equations and scheme does not come from the consistency conditions, but from the stability conditions. For hyperbolic schemes, we *do not* have (2.57) or (2.58).

Exercise 4.5 Show that the upwind scheme is first-order and the Lax-Wendroff scheme is second-order.

Proof of convergence . As usual, stability and consistency of the scheme, with sufficient regularity of the exact solution, are the ingredients to the convergence of the scheme. As in the parabolic case, we treat low and high wavenumbers separately. For low wavenumbers ($\Delta x|\xi| \leq \pi$), we have

$$\varepsilon^n(\xi) = (e^{-\Delta t P(i\xi)} - R_\Delta(\xi)) \sum_{k=0}^{n-1} e^{-(n-1-k)\Delta t P(i\xi)} R_\Delta^k(\xi) \hat{u}_0(\xi).$$

Here is the main difference with the parabolic case: $e^{-(n-1-k)\Delta t P(i\xi)}$ and $R_\Delta^k(\xi)$ are bounded but not small for high frequencies (notice that $|e^{-n\Delta t P(i\xi)}| = 1$ for $P(i\xi) = ia\xi$). So using the stability of the scheme, which states that $|R_\Delta^k(\xi)| \leq 1$, we deduce that

$$|\varepsilon^n(\xi)| \leq n|e^{-\Delta t P(i\xi)} - R_\Delta(\xi)||\hat{u}_0(\xi)| \leq C\Delta x^m (1 + |\xi|^{m+1})|\hat{u}_0(\xi)|,$$

using the order of convergence of the scheme.

For high wavenumbers $\Delta x|\xi| \geq \pi$, we deduce that

$$|\varepsilon^n(\xi)| \leq 2|\hat{u}_0(\xi)| \leq C\Delta x^m (1 + |\xi|^{m+1})|\hat{u}_0(\xi)|.$$

So the above inequality holds for all frequencies $\xi \in \mathbb{R}$. This implies that

$$\|\varepsilon^n(\xi)\| \leq C\Delta x^m \Big(\int_{\mathbb{R}} (1 + |\xi|^{2m+2})|\hat{u}_0(\xi)|^2 d\xi \Big)^{1/2}.$$

We have then proved the

Theorem 4.1 *Let us assume that $P(i\xi) = ia\xi$ and that $R_\Delta(\xi)$ is of order m, i.e. (4.7) holds. Assuming that the initial condition $u_0 \in H^{m+1}$, we obtain that*

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^m \|u_0\|_{H^{m+1}}, \quad \textit{for all } 0 \leq n \leq N = T/\Delta t. \tag{4.8}$$

The main difference with the parabolic case is that u_0 now needs to be in H^{m+1} (and not in H^m) to obtain an accuracy of order Δx^m. What if u_0 is not in H^m?

Exercise 4.6 Show that for any scheme of order m, we have

$$|e^{-i\Delta t P(i\xi)} - R_\Delta(\xi)| \leq C\Delta t \Delta x^{\alpha m} (1 + |\xi|^{\alpha(m+1)})$$

for all $0 \leq \alpha \leq 1$ and $|\xi|\Delta x \leq \pi$. Show that (4.8) is now replaced by

$$\|u(n\Delta t, x) - U^n(x)\| \leq C\Delta x^{\alpha m} \|u_0\|_{H^{\alpha(m+1)}}, \quad \text{for all } 0 \leq n \leq N = T/\Delta t. \tag{4.9}$$

Deduce that if u_0 is in H^s for $0 \leq s \leq m+1$, the error of convergence of the scheme will be of order $\frac{sm}{m+1}$.
We recall that piecewise constant functions are in the space H^s for all $s < 1/2$ (take $s = 1/2$ in the sequel to simplify). Deduce the order of the error of convergence for the upwind scheme and the Lax-Wendroff scheme. Verify this numerically.

5 Spectral methods

5.1 Unbounded grids and semidiscrete FT

The material of this section is borrowed from chapters 1 & 2 in Trefethen [2].

Finite differences were based on approximating

$$\frac{\partial}{\partial x}$$

by finite differences of the form

$$\partial_x, \qquad \overline{\partial}_x$$

for first-order approximations, or

$$\frac{1}{2}(\partial_x + \overline{\partial}_x),$$

for a second-order approximation, and so on for higher order approximations.

Spectral methods are based on using an approximation of very high order for the partial derivative. In matrix form, the second-order approximation given by (1.2) and fourth-order approximation given by (1.3) in [2] are replaced in the "spectral" approximation by the "infinite-order" matrix in (1.4).

How is this done? The main ingredient is "interpolation". We interpolate a set of values of a function at the points of a grid by a (smooth) function defined everywhere. We can then obviously differentiate this function and take the values of the differentiated function at the grid points. This is how derivatives are approximated in spectral methods and how (1.4) in [2] is obtained.

Let us be more specific and consider first the infinite grid $h\mathbb{Z}$ with grid points $x_j = jh$ for $j \in \mathbb{Z}$ and $h > 0$ given.

Let us assume that we are given the values v_j of a certain function $v(x)$ at the grid points, i.e. $v(x_j) = v_j$. What we want is to define a function $p(x)$ such that $p(x) = v_j$. This function $p(x)$ will then be an *interpolant* of the series $\{v_j\}$. To do so we define the *semidiscrete Fourier transform* as

$$\hat{v}(k) = h \sum_{j=-\infty}^{\infty} e^{-ikx_j} v_j, \qquad k \in [-\frac{\pi}{h}, \frac{\pi}{h}]. \tag{5.1}$$

It turns out that this transform can be inverted. We define the *inverse semidiscrete Fourier transform* as

$$v_j = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx_j} \hat{v}(k) dk. \tag{5.2}$$

The above definitions are nothing but the familiar Fourier series. Here, v_j are the Fourier coefficients of the periodic function $\hat{v}(k)$ on $[-\frac{\pi}{h}, \frac{\pi}{h}]$.

Exercise 5.1 Check that $\hat{v}(k)$ in (5.1) is indeed periodic of period $2\pi/h$.

It is then well known that the function $\hat{v}(k)$ can be reconstructed from its Fourier coefficients; this is (5.1).

Notice now that the inverse semidiscrete Fourier transform (a.k.a. Fourier series) in (5.2) is defined for every x_j of the grid $h\mathbb{Z}$, but could be defined more generally for every point $x \in \mathbb{R}$. This is how we define our interpolant, by simply replacing x_j in (5.2) by x:

$$\boxed{p(x) = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx} \hat{v}(k) dk.} \tag{5.3}$$

The function $p(x)$ is a very nice function. It can be shown that it is analytic (which means that the infinite Taylor expansion of $p(z)$ in the vicinity of every complex number z converges to p; so the function p is infinitely many times differentiable), and by definition it satisfies that $p(x_j) = v_j$.

Now assuming that we want to define w_j, an approximation of the derivative of v given by v_j at the grid points x_j. The spectral approximation simply consists of defining

$$w_j = p'(x_j), \tag{5.4}$$

where $p(x)$ is defined by (5.3). Notice that w_j a priori depends on all the values v_j. This is why the matrix (1.4) in [2] is infinite and full.

How do we calculate this matrix? Here we use the linearity of the differentiation. If

$$v_j = v_j^{(1)} + v_j^{(2)},$$

then we clearly have that

$$w_j = w_j^{(1)} + w_j^{(2)},$$

for the spectral differentials. So all we need to do is to consider the derivative of function v_j such that $v_n = 1$ for some $n \in \mathbb{Z}$ and $v_m = 0$ for $m \neq n$. We then obtain that

$$\hat{v}_n(k) = he^{-inkh},$$

and that

$$p_n(x) = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ik(x-nh)} dk = S_h(x - nh),$$

where $S_h(x)$ is the *sinc* function

$$S_h(x) = \frac{\sin(\pi x/h)}{\pi x/h}. \tag{5.5}$$

More generally, we obtain that the interpolant $p(x)$ is given by

$$\boxed{p(x) = \sum_{j=-\infty}^{\infty} v_j S_h(x - jh).} \tag{5.6}$$

An easy calculation shows that

$$S_h'(jh) = \begin{cases} 0, & j = 0 \\ \dfrac{(-1)^j}{jh}, & j \neq 0. \end{cases} \tag{5.7}$$

Exercise 5.2 Check this.

This implies that

$$p_n'(x_j) = \begin{cases} 0, & j = n \\ \dfrac{(-1)^{(j-n)}}{(j-n)h}, & j \neq n. \end{cases}$$

These are precisely the values of the entries (n, j) of the matrix in (1.4) in [2].

Notice that the same technique can be used to define approximations to derivatives of arbitrary order. Once we have the interpolant $p(x)$, we can differentiate it as many times as necessary. For instance the matrix corresponding to second-order differentiation is given in (2.14) of [2].

5.2 Periodic grids and Discrete FT

The simplest way to replace the infinite-dimensional matrices encountered in the preceding section is to assume that the initial function v is periodic. To simplify, it will be 2π periodic. The function is then represented by its value at the points $x_j = 2\pi j/N$ for some $N \in \mathbb{N}$, so that $h = 2\pi/N$. We also assume that N is even. Odd values of N are treated similarly, but the formulas are slightly different.

We know that periodic functions are represented by Fourier series. Here, the function is not only periodic, but also only given on the grid $(x_1, \cdots, x_N)$. So it turns out that its "Fourier series" is finite. This is how we define the *Discrete Fourier transform*

$$\hat{v}_k = h \sum_{j=1}^{N} e^{-ikx_j} v_j, \qquad k = -\frac{N}{2} + 1, \ldots, \frac{N}{2}. \tag{5.8}$$

The *Inverse Discrete Fourier Transform* is then given by

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx_j} \hat{v}_k, \qquad j = 1, \ldots, N. \tag{5.9}$$

The latter is recast as

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' e^{ikx_j} \hat{v}_k, \qquad j = 1, \ldots, N. \tag{5.10}$$

Here we define $\hat{v}_{-N/2} = \hat{v}_{N/2}$ and $\sum'$ means that the terms $k = \pm N/2$ are multiplied by $\frac{1}{2}$. The latter definition is more symmetrical than the former, although both are obviously equivalent. The reason we introduce (5.10) is that the two definitions do not yield the same interpolant, and the interpolant based on (5.10) is more symmetrical.

As we did for infinite grids, we can extend (5.10) to arbitrary values of x and not only the values on the grid. The *discrete interpolant* is now defined by

$$\boxed{p(x) = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' e^{ikx} \hat{v}_k, \qquad x \in [0, 2\pi].} \tag{5.11}$$

Notice that this interpolant is obviously 2π periodic, and is a *trigonometric polynomial* of order (at most) $N/2$.

We can now easily obtain a spectral approximation of $v(x)$ at the grid points: we simply differentiate $p(x)$ and take the values of the solution at the grid points. Again, this corresponds to

$$w_j = p'(x_j), \qquad j = 1, \ldots, N.$$

Exercise 5.3 Calculate the $N \times N$ matrix D_N which maps the values v_j of the function $v(x)$ to the values w_j of the function $p'(x_j)$. The result of the calculation is given on p.21 in [2].

5.3 Fast Fourier Transform (FFT)

In the preceding subsections, we have seen how to calculate spectral derivatives at grid points w_j of a function defined by v_j at the same gridpoints. We have also introduced matrices that map v_j to w_j. Our derivation was done in the physical domain, in the sense that we have constructed the interpolant $p(x)$, taken its derivative, p' and evaluated the derivative $p'(x_j)$ at the grid points. Another way to consider the derivation is to see what it means to "take a derivative" in the Fourier domain. This is what FFT is based on.

Remember that (classical) Fourier transforms replace derivation by multiplication by $i\xi$; see (2.27). The interpolations of the semidiscrete and discrete Fourier transforms have very similar properties. For instance we deduce from (5.3) that

$$p'(x) = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx} ik\hat{v}(k) dk.$$

So again, in the Fourier domain, differentiation really means replacing $\hat{v}(k)$ by $ik\hat{v}(k)$. For the discrete Fourier transform, we deduce from (5.11) that

$$p'(x) = \frac{1}{2\pi} {\sum_{k=-N/2}^{N/2}}' e^{ikx} ik\hat{v}_k.$$

Notice that by definition, we have $\hat{v}_{-N/2} = \hat{v}_{N/2}$ so that $i(-N/2)\hat{v}_{-N/2} + iN/2\hat{v}_{N/2} = 0$. In other words, we have

$$p'(x) = \frac{1}{2\pi} {\sum_{k=-N/2+1}^{N/2-1}}' e^{ikx} ik\hat{v}_k.$$

This means that the Fourier coefficients $\hat{v}_k$ are indeed replaced by $ik\hat{v}_k$ when we differentiate $p(x)$, expect for the value of $k = N/2$, for which the Fourier coefficient vanishes.

So, to calculate w_j from v_j in the periodic case, we have the following procedure

1. Calculate $\hat{v}_k$ from v_j for $-N/2 + 1 \leq k \leq N/2$.
2. Define $\hat{w}_k = ik\hat{v}_k$ for $-N/2 + 1 \leq k \leq N/2 - 1$ and $\hat{w}_{N/2} = 0$.
3. Calculate w_j from $\hat{w}_j$ by Inverse Discrete Fourier transform.

Said in other words, the above procedure is the discrete analog of the formula

$$f'(x) = \mathcal{F}^{-1}_{\xi\to x}\Big[i\xi\mathcal{F}_{x\to\xi}[f(x)]\Big].$$

On a computer, the multiplication by ik (step 2) is fast. The main computational difficulties come from the implementation of the discrete Fourier transform and its inversion (steps 1 and 3). A very efficient implementation of steps 1 and 3 is called the FFT, the Fast Fourier Transform. It is an algorithm that performs steps 1 and 3 in $O(N\log N)$ operations. This has to be compared with a number of operations of order $O(N^2)$ if the formulas (5.8) and (5.9) are being used.

Exercise 5.4 Check the order $O(N^2)$.

We are not going to describe how FFT works; let us merely state that it is a very convenient alternative to constructing the matrix D_N to calculate the spectral derivative. We refer to [2] Programs 4 and 5 for details.

5.4 Spectral approximation for unbounded grids

We have introduced several interpolants to obtain approximations of the derivatives of functions defined by their values on a (discrete) grid. It remains to know what sort of error one makes by introducing such interpolants. Chapter 4 in [2] gives some answer to this question. Here we shall give a slightly different answer based on the Sobolev scale introduced in the analysis of finite differences.

Let $u(x)$ be a function defined for $x\in\mathbb{R}$ and define $v_j = u(x_j)$. The Fourier transform of $u(x)$ is denoted by $\hat{u}(k)$. We then define the semidiscrete Fourier transform $\hat{v}(k)$ by (5.1) and the interpolant $p(x)$ by (5.3).

The question is then: what is the error between $u(x)$ and $p(x)$? The answer relies on one thing: how regular is $u(x)$. The analysis of the error is based on estimates in the Fourier domain. Smooth functions $u(x)$ correspond to functions $\hat{u}(k)$ that decay fast in k.

The first theoretical ingredient, which is important in its own right, is the *aliasing formula*, also called the *Poisson summation formula*.

Theorem 5.1 *Let $u\in L^2(\mathbb{R})$ be sufficiently smooth. Then for all $k\in[-\pi/h,\pi/h]$,*

$$\hat{v}(k) = \sum_{j=-\infty}^{\infty}\hat{u}\Big(k+\frac{2\pi j}{h}\Big). \tag{5.12}$$

In other words the aliasing formula relates the discrete Fourier transform $\hat{v}(k)$ to the continuous Fourier transform $\hat{u}(k)$.

Derivation of the aliasing formula. We write

$$\begin{aligned}\hat{v}(k) &= h\sum_{j=-\infty}^{\infty} e^{-ikhj}v_j \\ &= \sum_{j=-\infty}^{\infty}\int_{\mathbb{R}} \hat{u}(k')e^{ihj(k'-k)}\frac{hdk'}{2\pi} \\ &= \int_{\mathbb{R}} \hat{u}(k+\frac{2\pi l}{h})\sum_{j=-\infty}^{\infty} e^{i2\pi lj}dl,\end{aligned}$$

using the change of variables $h(k'-k) = 2\pi l$ so that $hdk' = 2\pi dl$. We will show that

$$\sum_{j=-\infty}^{\infty} e^{i2\pi lj} = \sum_{j=-\infty}^{\infty}\delta(l-j). \tag{5.13}$$

Assuming that this equality holds, we then obtain that

$$\hat{v}(k) = \int_{\mathbb{R}} \hat{u}(k+\frac{2\pi l}{h})\sum_{j=-\infty}^{\infty}\delta(l-j)dl = \sum_{j=-\infty}^{\infty}\hat{u}(k+\frac{2\pi j}{h}),$$

which is what was to be proved. It thus remains to show (5.13). Here is a simple explanation. Take the Fourier series of a $1-$periodic function $f(x)$

$$c_n = \int_{-1/2}^{1/2} e^{-i2\pi nx}f(x)dx.$$

Then the inversion is given by

$$f(x) = \sum_{n=-\infty}^{\infty} e^{i2\pi nx}c_n.$$

Now take $f(x)$ to be the delta function $f(x) = \delta(x)$. We then deduce that we have $c_n = 1$ and for $x \in (0,1)$,

$$\sum_{n=-\infty}^{\infty} e^{i2\pi nx} = \delta(x).$$

Let us now extend both sides of the above relation by periodicity to the whole line $\mathbb{R}$. The left-hand side is a 1-periodic in x and does not change. The right hand side now takes the form

$$\sum_{j=-\infty}^{\infty}\delta(x-j).$$

This is nothing but the periodic delta function. This proves (5.13). □

Now that we have the aliasing formula, we can analyze the difference $u(x) - p(x)$. We'll do it in the L^2 sense, since

$$\|u(x)-p(x)\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}}\|\hat{u}(k)-\hat{p}(k)\|_{L^2(\mathbb{R})}. \tag{5.14}$$

We now show the following result:

Theorem 5.2 *Let us assume that $u(x) \in H^m(\mathbb{R})$ for $m > 1/2$. Then we have*

$$\|u(x) - p(x)\|_{L^2(\mathbb{R})} \leq Ch^m \|u\|_{H^m(\mathbb{R})}, \tag{5.15}$$

where the constant C is independent of h and u.

Proof. We first realize that

$$\hat{p}(k) = \chi(k)\hat{v}(k)$$

by construction, where $\chi(k) = 1$ if $k \in [-\pi/h, \pi/h]$ and $\chi(k) = 0$ otherwise is the *characteristic function* of the interval $[-\pi/h, \pi/h]$. In other words, $p(x)$ is *band-limited*, i.e. has no high frequencies. We thus deduce that

$$\int_{\mathbb{R}} |\hat{u}(k) - \hat{p}(k)|^2 dk = \int_{\mathbb{R}} (1 - \chi(k))|\hat{u}(k)|^2 dk + \int_{-\pi/h}^{\pi/h} \Big| \sum_{j \in \mathbb{Z}^*} \hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|^2 dk,$$

using (5.12). We have denoted by $\mathbb{Z}^* = \mathbb{Z}\backslash\{0\}$. Looking carefully at the above terms, one realizes that the above error only involves $\hat{u}(k)$ for values of k outside the interval $[-\pi/h, \pi/h]$. In other words, if $\hat{u}(k)$ has compact support inside $[-\pi/h, \pi/h]$, then $\hat{p}(k) = \hat{u}(k)$, hence $p(x) = u(x)$ and the interpolant is *exact*. When $\hat{u}(k)$ does not vanish outside $[-\pi/h, \pi/h]$, i.e. when the high frequency content is present, we have to use the regularity of the function $u(x)$ to control the above term. This is what we now do. The first term is easy to deal with. Indeed we have

$$\int_{\mathbb{R}} (1 - \chi(k))|\hat{u}(k)|^2 dk \leq h^{2m} \int_{\mathbb{R}} (1 + k^{2m})|\hat{u}(k)|^2 dk \leq h^{2m} \|u\|^2_{H^m(\mathbb{R})}.$$

This is simply based on realizing that $|hk| \geq \pi$ when $\chi(k) = 1$. The second term is a little more painful. The Cauchy-Schwarz inequality tells us that

$$\Big| \sum_j a_j b_j \Big|^2 \leq \sum_j |a_j|^2 \sum_j |b_j|^2.$$

This is a mere generalization of $|x \cdot y| \leq |x||y|$. Let us choose

$$a_j = \Big|\hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|\Big(1 + |k + \frac{2\pi j}{h}|\Big)^m, \qquad b_j = \Big(1 + |k + \frac{2\pi j}{h}|\Big)^{-m}$$

We thus obtain that

$$\Big| \sum_{j \in \mathbb{Z}^*} \hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|^2 \leq \sum_{j \in \mathbb{Z}^*} \Big|\hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|^2 \Big(1 + |k + \frac{2\pi j}{h}|\Big)^{2m} \sum_{j \in \mathbb{Z}^*} \Big(1 + |k + \frac{2\pi j}{h}|\Big)^{-2m}.$$

Notice that

$$\sum_{j \in \mathbb{Z}^*} \Big(1 + |k + \frac{2\pi j}{h}|\Big)^{-2m} \leq h^{2m} \sum_{j \in \mathbb{Z}^*} \frac{1}{(2\pi(1 + |j|))^{2m}}, \qquad \text{for } |k| \leq \pi/h.$$

For $m > 1/2$, the above sum is convergent so that

$$\sum_{j \in \mathbb{Z}^*} \Big(1 + |k + \frac{2\pi j}{h}|\Big)^{-2m} \leq Ch^{2m},$$

where C depends on m but not on u or h. This implies that

$$\int_{-\pi/h}^{\pi/h} \Big| \sum_{j\in\mathbb{Z}^*} \hat{u}\Big(k+\frac{2\pi j}{h}\Big)\Big|^2 dk \le Ch^{2m} \int_{-\pi/h}^{\pi/h} \sum_{j\in\mathbb{Z}^*} \Big|\hat{u}\Big(k+\frac{2\pi j}{h}\Big)\Big|^2 \Big(1+|k+\frac{2\pi j}{h}|\Big)^{2m}$$
$$\le Ch^{2m} \int_{\mathbb{R}} |\hat{u}(k)|^2 (1+|k|)^{2m} dk = Ch^{2m}\|u\|^2_{H^m}.$$

This concludes the proof of the theorem. □

Let us emphasize something we saw in the course of the above proof. When $u(x)$ is band-limited so that $\hat{u}(k) = 0$ outside $[-\pi/h, \pi/h]$, we have that the interpolant is exact; namely $p(x) = u(x)$. However, we have seen that the interpolant is determined by the values $v_j = u(x_j)$.

More precisely, we saw that

$$p(x) = \sum_{n=-\infty}^{\infty} v_n S_h(x - nh),$$

where S_h is the sinc function defined in (5.5). This proves the extremely important *Shannon-Whittaker sampling theorem*:

Theorem 5.3 *Let us assume that the function $u(x)$ is such that $\hat{u}(k) = 0$ for k outside $[-\pi/h, \pi/h]$. Then $u(x)$ is completely determined by its values $v_j = u(hj)$ for $j \in \mathbb{Z}$. More precisely, we have*

$$u(x) = \sum_{n=-\infty}^{\infty} u(hj) S_h(x - jh), \qquad S_h(x) = \frac{\sin(\pi x/h)}{\pi x/h}. \tag{5.16}$$

This result has quite important applications in communication theory. Assuming that a signal has no high frequencies (such as in speech, since the human ear cannot detect frequencies above 25kHz), there is an adapted sampling value h such that the signal can be reconstructed from the values it takes on the grid $h\mathbb{Z}$. This means that a continuous signal can be compressed onto a discrete grid with no error. When h is chosen too large so that frequencies above π/h are present in the signal, then the interpolant is not exact, and the compression in the signal has some error. This error is precisely called *aliasing* error and is quantified by Theorem 5.1.

Let us finish this section by an analysis of the error between the derivatives of $u(x)$ and those of $p(x)$. Remember that our initial objective is to obtain an approximation of $u'(x_j)$ by using $p'(x_j)$. Minor modifications in the proof of Theorem 5.2 yield the following result:

Theorem 5.4 *Let us assume that $u(x) \in H^m(\mathbb{R})$ for $m > 1/2$ and let $n \le m$. Then we have*

$$\|u^{(n)}(x) - p^{(n)}(x)\|_{L^2(\mathbb{R})} \le Ch^{(m-n)}\|u\|_{H^m(\mathbb{R})}, \tag{5.17}$$

where the constant C is independent of h and u.

The proof of this theorem consists of realizing that

$$\|u^{(n)}(x) - p^{(n)}(x)\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}} \|(ik)^n(\hat{u}(k) - \hat{p}(k))\|_{L^2(\mathbb{R})}.$$

Moreover we have

$$\int_{\mathbb{R}} |k|^{2n}|\hat{u}(k) - \hat{p}(k)|^2 dk = \int_{\mathbb{R}} |k|^{2n}(1-\chi(k))|\hat{u}(k)|^2 dk$$
$$+ \int_{-\pi/h}^{\pi/h} |k|^{2n} \Big| \sum_{j\in\mathbb{Z}^*} \hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|^2 dk.$$

The first term on the right hand side is bounded by

$$h^{2(m-n)} \int (1+k^{2m})|\hat{u}(k)|^2 dk = \Big(h^{m-n}\|u\|_{H^m}\Big)^2.$$

The second term is bounded by

$$h^{-2n} \int_{-\pi/h}^{\pi/h} \Big| \sum_{j\in\mathbb{Z}^*} \hat{u}\Big(k + \frac{2\pi j}{h}\Big)\Big|^2 dk,$$

which is bounded, as we saw in the proof of Theorem 5.2 by $h^{-2n}h^{2m}\|u\|^2_{H^m}$. This proves the theorem. □

This result is important for us: it shows that the error one makes by approximating a $n-$th order derivative of $u(x)$ by using $p^{(n)}(x)$ is of order h^{m-n}, where m is the H^m-regularity of u. So when u is very smooth, $p^{(n)}(x)$ converges extremely fast to $u^{(n)}(x)$ as $N \to \infty$. This is the greatest advantage of spectral methods.

5.5 Application to PDE's

An elliptic equation. Consider the simplest of elliptic equations

$$-u''(x) + \sigma u(x) = f(x), \qquad x \in \mathbb{R}, \tag{5.18}$$

where the absorption $\sigma > 0$ is a constant positive parameter. Define $\alpha = \sqrt{\sigma}$. We verify that

$$G(x) = \frac{1}{2\alpha} e^{-\alpha|x|}, \tag{5.19}$$

is the Green function of the above problem, whose solution is thus

$$u(x) = \int_{\mathbb{R}} G(x-y)f(y)dy = \frac{1}{2\alpha}\int_{\mathbb{R}} e^{-\alpha|x-y|} f(y)dy. \tag{5.20}$$

Note that this may also be obtained in the Fourier domain, where

$$\hat{u}(\xi) = \frac{\hat{f}(\xi)}{\sigma + \xi^2}. \tag{5.21}$$

Discretizing the above equation using the second-order finite difference method:

$$-\frac{U(x+h) + U(x-h) - 2U(x)}{h^2} + U(x) = f(x), \tag{5.22}$$

we obtain as in (3.19) in section 3 an error estimate of the form

$$\|u - U\|_{L^2} \le C\ h^2\ \|f\|_{H^2}. \tag{5.23}$$

Let us now compare this result to the discretization based on the spectral method. Spectral methods are based on replacing functions by their spectral interpolants. Let $p(x)$ be the spectral interpolant of $u(x)$ and $F(x)$ the spectral interpolant of the source term $f(x)$. Now $p(x)$ and $F(x)$ only depend on the values they take at hj for $j \in \mathbb{Z}$. We may now calculate the second-order derivative of $p(x)$ (see the following paragraph for an explicit expression). The equation we now use to deduce $p(x)$ from $F(x)$ is the equation (5.18):

$$-p''(x) + \sigma p(x) = F(x), \qquad x \in \mathbb{R}.$$

Note that this problem may be solved by inverting a matrix of the form $-N^2 + \sigma I$, where N is the matrix of differentiation. The error $u - p$ now solves the equation

$$-(u-p)''(x) + \sigma(u-p) = (f-F)(x).$$

Looking at this in the Fourier domain (and coming back) we easily deduce that

$$\|u-p\|_{H^2(\mathbb{R})} \leq C\|f-F\|_{L^2(\mathbb{R})} \leq Ch^m \|f\|_{H^m(\mathbb{R})}, \tag{5.24}$$

where the latter equality comes from Theorem 5.2. Here m is arbitrary provided that f is arbitrarily smooth. We thus obtain that the spectral method has an "infinite" order of accuracy, unlike finite difference methods. When the source terms and the solutions of the PDEs we are interested in are smooth (spatially), the convergence properties of spectral methods are much better than those of finite difference methods.

A parabolic equation. Let us now apply the above theory to the computation of solutions of the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0,$$

on the real line $\mathbb{R}$. We introduce U_j^n the approximate solution at time $T_n = n\Delta t$, $1 \leq n \leq N$ and point $x_j = hj$ for $j \in \mathbb{Z}$.

Let us consider the Euler explicit scheme based on a spectral approximation of the spatial derivatives. The interpolant at time T_n is given by

$$p^n(x) = \sum_{j=-\infty}^{\infty} U_j^n S_h(x - jh),$$

so that the approximation to the second derivative is given by

$$(p^n)''(kh) = \sum_{j=-\infty}^{\infty} S_h''((k-j)h) U_j^n = \sum_{j=-\infty}^{\infty} S_h''(jh) U_{k-j}^n.$$

So the Euler explicit scheme is given by

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \sum_{k=-\infty}^{\infty} S_h''(kh) U_{j-k}^n. \tag{5.25}$$

It now remains to see how this scheme behaves. We shall only consider stability here (i.e. see how Δt and h have to be chosen so that the discrete solution does not blow

up as $n \to T/\Delta t$). As far as convergence is concerned, the scheme will be first order in time (accuracy of order Δt) and will be of spectral accuracy (infinite order) provided that the initial solution is sufficiently smooth. This is plausible from the analysis of the error made by replacing a function by its interpolant that we made earlier.

Let us come back to the stability issue. As we did for finite differences, we realize that we can look at a scheme defined for all $x \in \mathbb{R}$ and not only $x = x_j$:

$$\frac{U^{n+1}(x) - U^n(x)}{\Delta t} = \sum_{k=-\infty}^{\infty} S_h''(kh) U^n(x - hk). \tag{5.26}$$

Once we have realized this, we can pass to the Fourier domain and obtain that

$$\hat{U}^{n+1}(\xi) = \Big[1 + \Delta t \sum_{k=-\infty}^{\infty} S_h''(kh) e^{-ikh\xi}\Big] \hat{U}^n(\xi). \tag{5.27}$$

So the stability of the scheme boils down to making sure that

$$R(\xi) = 1 + \Delta t \sum_{k=-\infty}^{\infty} S_h''(kh) e^{-ikh\xi}, \tag{5.28}$$

stays between -1 and 1 for all frequencies ξ.

The analysis is more complicated than for finite differences because of the infinite sum that appears in the definition of R. Here is a way to analyze $R(\xi)$. We first realize that

$$S_h''(kh) = \begin{cases} \dfrac{-\pi^2}{3h^2} & k = 0 \\ 2\dfrac{(-1)^{k+1}}{h^2 k^2} & k \neq 0. \end{cases}$$

Exercise 5.5 Check this.

So we can recast $R(\xi)$ as

$$R(\xi) = 1 - \Delta t \frac{\pi^2}{3h^2} - \frac{2\Delta t}{h^2} \sum_{k \neq 0} \frac{e^{i(h\xi - \pi)k}}{k^2}.$$

Exercise 5.6 Check this.

We now need to understand the remaining infinite sum. It turns out that we can have access to an exact formula. It is based on using the Poisson formula

$$\sum_j e^{i2\pi l j} = \sum_j \delta(l - j),$$

which we recast for $l \in [-1/2, 1/2]$ as

$$\sum_{j \neq 0} e^{i2\pi l j} = \delta(l) - 1.$$

Notice that both terms have zero mean on $[-1/2, 1/2]$ (i.e. there is no 0 frequency). We now integrate these terms in l. We obtain that

$$\sum_{j\neq 0} \frac{e^{i2\pi lj}}{i2\pi j} = H(l) - l - d_0,$$

where $H(l)$ is the Heaviside function, such that $H(l) = 1$ for $l > 0$ and $H(l) = 0$ for $l < 0$. The constant d_0 is now fixed by making sure that the above right-hand side remains 1-periodic (as is the left-hand side). Another way of looking at it is to make sure that both sides have zero mean on $(-1/2, 1/2)$. We thus find that $d_0 = 1/2$.

Exercise 5.7 Check this.

Let us integrate one more time. We now obtain that

$$\sum_{j\neq 0} \frac{e^{i2\pi lj}}{-4\pi^2 j^2} = lH(l) - \frac{l^2}{2} - \frac{l}{2} - d_1,$$

where d_1 is again chosen so that the right-hand side has zero mean. We obtain that $d_1 = 1/12$.

Exercise 5.8 Check this.

To sum up our calculation, we have obtained that

$$\sum_{j\neq 0} \frac{e^{i2\pi lj}}{j^2} = 4\pi^2\Big(\frac{l^2}{2} + \frac{1}{12} - l\Big(H(l) - \frac{1}{2}\Big)\Big).$$

This shows that

$$R(\xi) = 1 - \frac{\Delta t}{h^2}\Big(\frac{\pi^2}{3} + 8\pi^2\Big(\frac{l^2}{2} + \frac{1}{12} - l\Big(H(l) - \frac{1}{2}\Big)\Big)\Big), \qquad l = \frac{h\xi}{2\pi} - \frac{1}{2}. \tag{5.29}$$

A plot of the function $l \to l^2/2 + 1/12 - l(H(l) - 1/2)$ is given in Fig. 5.1. Notice that this function is 1-periodic and continuous. This is justified as the delta function is concentrated at one point, its integral is a piecewise linear function $(H(l) - 1/2 - l)$ and is discontinuous, and its second integral is continuous (and is piecewise quadratic).

Exercise 5.9 Check that $l(H(l) - \frac{1}{2}) = \frac{1}{2}|l|$. Deduce that

$$R(\xi) = 1 - \frac{\Delta t}{h^2}\pi^2\Big(1 + 4l(1 - |l|)\Big),$$

so that

$$R(\xi) = 1 - \frac{\Delta t}{h^2}h^2\xi^2 := 1 - \Delta t\xi^2, \qquad 0 \leq |\xi| \leq \frac{\pi}{h}.$$

[Check the above for $h\xi \leq \pi$ and use (5.28) to show that $R(\xi) = R(-\xi)$ and that $R(\xi)$ is $\frac{2\pi}{h}$-periodic.]

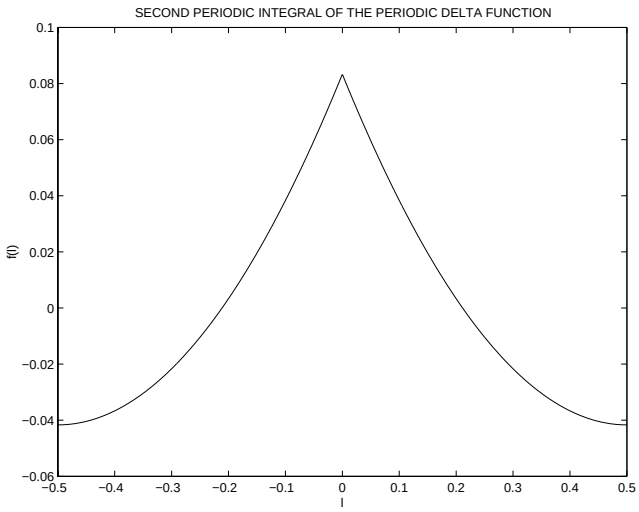

Figure 5.1: Plot of the function $l \to f(l) = l^2/2 + 1/12 - l(H(l) - 1/2)$. The maximal value is $1/12$ and the minimal value $-1/24$.

Since $R(\xi)$ is a periodized version of the symbol $1 - \Delta t \xi^2$, we deduce that

$$1 - \pi^2 \frac{\Delta t}{h^2} \leq R(\xi) \leq 1,$$

and that the minimal value is reached for $l = 0$, i.e. for frequencies

$$\xi = \frac{2\pi}{h}(m + \frac{1}{2}), \qquad m \in \mathbb{Z},$$

and that the maximum is reached for $l = \pm 1/2$, i.e. for frequencies

$$\xi = \frac{2\pi m}{h}, \qquad m \in \mathbb{Z}.$$

Notice that $R(0) = 0$ as it should, and that $R(\pi/h) = 1 - \pi^2 \Delta t/h^2$, which is a quite bad approximation of $e^{-\Delta t \pi^2/(h^2)}$ for the exact solution. But as usual, we do not expect frequencies of order h^{-1} to be well approximated on a grid of size h.

Since we need $|R(\xi)| \leq 1$ for all frequencies to obtain a stable scheme, we see that

$$\lambda = \frac{\Delta t}{h^2} \leq \frac{2}{\pi^2}, \tag{5.30}$$

is necessary to obtain a convergent scheme as $h \to 0$ with λ fixed.

6 Introduction to finite element methods

We now turn to one of the most useful methods in numerical simulations: the finite element method. This will be very introductory. The material mostly follows the presentation in [1].

6.1 Elliptic equations

Let Ω be a convex open domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. We will assume that $\partial\Omega$ is either sufficiently smooth or a polygon. On Ω we consider the boundary value problem

$$\begin{aligned} &-\Delta u(\mathbf{x}) + \sigma(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2) \in \Omega \\ &u(\mathbf{x}) = 0, \qquad \mathbf{x} \in \partial\Omega. \end{aligned} \tag{6.1}$$

The Laplace operator Δ is defined as

$$\Delta = \frac{\partial^2}{\partial {x_1}^2} + \frac{\partial^2}{\partial {x_2}^2}.$$

We assume that the *absorption* coefficient $\sigma(\mathbf{x}) \in L^\infty(\Omega)$ satisfies the constraint

$$\sigma(\mathbf{x}) \geq \sigma_0 > 0, \qquad \mathbf{x} \in \bar{\Omega}. \tag{6.2}$$

We also assume that the source term $f \in L^2(\Omega)$ (i.e., is square integrable).

This subsection is devoted to showing the existence of a solution to the above problem and recalling some elementary properties that will be useful in the analysis of discretized problems.

Let us first assume that there exists a sufficiently smooth solution $u(\mathbf{x})$ to (6.1). Let $v(\mathbf{x})$ be a sufficiently smooth *test function* defined on Ω and such that $v(\mathbf{x}) = 0$ for $\mathbf{x} \in \partial\Omega$. Upon multiplying (6.1) by $v(\mathbf{x})$ and integrating over Ω, we obtain by integrations by parts using Green's formula that:

$$a(u, v) := \int_\Omega \big(\nabla u \cdot \nabla v + \sigma uv\big) d\mathbf{x} = \int_\Omega f v d\mathbf{x} := L(v). \tag{6.3}$$

Exercise 6.1 Prove the above formula.

Let us introduce a few Hilbert spaces. For $k \in \mathbb{N}$, we define

$$H^k(\Omega) \equiv H^k = \Big\{ v \in L^2(\Omega); \quad D^\alpha v \in L^2(\Omega), \quad \text{for all } |\alpha| \leq k \}. \tag{6.4}$$

We recall that for a multi-index $\alpha = (\alpha_1, \alpha_2)$ (in two space dimensions) for $\alpha_k \in \mathbb{N}$, $1 \leq k \leq 2$, we denote $|\alpha| = \alpha_1 + \alpha_2$ and

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}}.$$

Thus $v \in H^k \equiv H^k(\Omega)$ if all its partial derivatives of order up to k are square integrable in Ω. The Hilbert space is equipped with the *norm*

$$\|u\|_{H^k} = \Big(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^2}^2 \Big)^{1/2}. \tag{6.5}$$

We also define the *semi-norm* on H^k:

$$|u|_{H^k} = \Big(\sum_{|\alpha|=k} \|D^\alpha u\|_{L^2}^2 \Big)^{1/2}. \tag{6.6}$$

Note that the above semi-norm is not a norm, for all polynomials p_{k-1} of order at most $k-1$ are such that $|p_{k-1}|_{H^k} = 0$.

Because of the choice of boundary conditions in (6.1), we also need to introduce the Hilbert space

$$H_0^1(\Omega) = \Big\{ v \in H^1(\Omega); \quad v_{|\partial\Omega} = 0 \Big\}. \tag{6.7}$$

We can show that the above space is a closed subspace of $H^1(\Omega)$, where both spaces are equipped with the norm $\|\cdot\|_{H^1}$. Showing this is not completely trivial and requires understanding the operator that restricts a function defined on Ω to the values it takes at the boundary $\partial\Omega$. The restriction is defined for sufficiently smooth functions only (being an element in H^1 is sufficient, not necessary). For instance: "the restriction to $\partial\Omega$ of a function in $L^2(\Omega)$" means nothing. This is because L^2 functions are defined up to modifications on sets of measure zero (for the Lebesgue measure) and that $\partial\Omega$ is precisely a set of measure zero.

We can show that (6.3) holds for any function $v \in H_0^1(\Omega)$. This allows us to "replace" the problem (6.1) by the new problem in *variational form*:

Find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$, we have

$$a(u, v) = L(v). \tag{6.8}$$

Theorem 6.1 *There is a unique solution to the above problem* (6.8).

Proof. The theorem is a consequence of the Riesz representation theorem. The main ingredients of the proof are the following. First, note that $a(u, v)$ is a *symmetric* bilinear form on the Hilbert space $V = H_0^1(\Omega)$. It is clearly linear in u and v and one verifies that $a(u, v) = a(v, u)$.

The bilinear form $a(u, v)$ is also positive definite, in the sense that $a(v, v) > 0$ for all $v \in V$ such that $v \neq 0$. Indeed, $a(u, v)$ is even better than this: it is *coercive* in V, in the sense that there exists $\alpha > 0$ such that

$$a(u, u) \geq \alpha \|u\|_V^2, \qquad \text{for all } v \in V. \tag{6.9}$$

Indeed, from the definition of $a(u, v)$ and the constraint on σ, we can choose $\alpha = \min(1, \sigma_0)$.

Since $a(u, v)$ is a symmetric, positive definite, bilinear form on V, it defines an *inner product* on V (by definition of an inner product). The norm associated to this inner product is given by

$$\|u\|_a = a^{1/2}(u, u), \qquad u \in V. \tag{6.10}$$

The Riesz representation theorem then precisely states the following: Let V a Hilbert space with an inner product $a(\cdot, \cdot)$. Then for each bounded linear form L on V, there is a unique $u \in V$ such that $L(v) = a(u, v)$ for all $v \in V$. Theorem 6.1 is thus equivalent

to showing that $L(v)$ is bounded on V, which means the existence of a constant $\|L\|_{V^*}$ such that $|L(v)| \leq \|L\|_{V^*} a^{1/2}(v,v)$ for all $v \in V$. However we have

$$|L(v)| \leq \|f\|_{L^2}\|v\|_{L^2} = \frac{\|f\|_{L^2}}{\sqrt{\sigma_0}}(\sqrt{\sigma_0}\|v\|_{L^2}) \leq \frac{\|f\|_{L^2}}{\sqrt{\sigma_0}} a^{1/2}(v,v).$$

This yields the bound for L and the existence of a solution to (6.8). Now assume that u and w are solutions. Then by taking differences in (6.8), we deduce that $a(u-w,v) = 0$ for all $v \in V$. Choose now $v = u - w$ to get that $a(u-w, u-w) = 0$. However, since a is positive definite, this implies that $u - w = 0$ and the uniqueness of the solution. □

Remark 6.2 The solution to the above problem admits another interpretation. Indeed let us define the functional

$$J(u) = \frac{1}{2}a(u,u) - L(u). \tag{6.11}$$

We can then show that u is a solution to (6.8) is equivalent to u being the minimum of $J(u)$ on $V = H_0^1$.

Remark 6.3 Theorem 6.1 admits a very simple proof based on the Riesz representation theorem. However it requires $a(u,v)$ to be symmetric, which may not be the case in practice. The Lax-Milgram theory is then the tool to use. What it says is that provided that $a(u,v)$ is coercive and bounded on a Hilbert space V (equipped with norm $\|\cdot\|_V$), in the sense that there are two constants C_1 and α such that

$$\alpha\|v\|_V^2 \leq a(v,v), \qquad |a(v,w)| \leq C_1\|v\|_V\|w\|_V, \quad \text{for all } v, w \in V, \tag{6.12}$$

and provided that the functional $L(v)$ is bounded on V, in the sense that there exists a constant $\|L\|_{V^*}$ such that

$$|L(v)| \leq \|L\|_{V^*}\|v\|_V, \quad \text{for all } v \in V$$

then the abstract problem

$$a(u,v) = L(v), \qquad \text{for all } v \in V, \tag{6.13}$$

admits a unique solution $u \in V$. This abstract theory is extremely useful in many practical problems.

Regularity issues. The above theory gives us a *weak solution* $u \in H_0^1$ to the problem (6.1). The solution is "weak" because the elliptic problem should be considered in it variational form (6.3) rather than its PDE form (6.1). The Laplacian of a function in H_0^1 is not necessarily a function. So u is not a *strong solution* (which is defined as a C^2 solution of (6.1)).

It thus remains to understand whether the regularity of $u \in H_0^1$ is optimal. The answer is no in most cases. What we can show is that for smooth boundaries $\partial\Omega$ and for convex polygons $\partial\Omega$, we have

$$\|u\|_{H^2} \leq C\|f\|_{L^2}. \tag{6.14}$$

When $f \in L^2$, we thus deduce that $u \in H^2 \cap H_0^1$. Deriving such results is beyond the scope of these notes and we will simply admit them. The main aspect of this regularity result is that u has two more degrees of regularity than f; when $f \in L^2$, the second-order partial derivatives of u are also in L^2. This may be generalized as follows. For sufficiently smooth $\partial\Omega$, or for very special polygons Γ (such as rectangles), we have the following result:

$$\|u\|_{H^{k+2}} \leq C\|f\|_{H^k}, \qquad k \geq 0. \tag{6.15}$$

This means that the solution u may be quite regular provided that the source term f is sufficiently regular. Moreover, a classical Sobolev inequality theorem says that a function $u \in H^k$ for $k > d/2$, where d is space dimension, is also of class C^0. In dimension $d = 2$, this means that as soon as $f \in H^k$ for $k > 1$, then u is of class C^2, i.e., is a strong solution of (6.1).

As we saw for finite difference and spectral methods the regularity of the solution dictates the accuracy of the discretized solution. The above regularity results will thus be of crucial importance in the analysis of finite element methods which we take up now.

6.2 Galerkin approximation

The above theory of existence is fundamentally different from what we used in the finite difference and spectral methods: we no longer have (6.1) to discretize but rather (6.3). The main difficulty in solving (6.3) is that $V = H_0^1$ is infinite dimensional. This however gives us the right idea to discretize (6.3): simply replace V by a discrete approximation V_h and solve (6.3) in V_h rather than V. We then have to choose the family of spaces V_h for some parameter $h \to 0$ (think of h as a mesh size) such that the solution converges to u in a reasonable sense.

The Galerkin method is based on choosing finite dimensional subspaces $V_h \subset V$. In our problem (6.1), this means choosing V_h as subsets of H_0^1. This requires some care: functions in H_0^1 have derivatives in L^2. So piecewise constant functions for instance, are not elements of H_0^1 (because they can jump and the derivatives of jumps are too singular to be elements of L^2), and V_h can therefore *not* be constructed using piecewise constant functions.

Let us assume that we have a sequence of finite dimensional subspaces $V_h \subset V$. Then the spaces V_h are Hilbert spaces. Assuming that the bilinear form $a(u, v)$ is coercive and bounded on V (see Rem. 6.3), then clearly it remains coercive and bounded on V_h (because it is specifically chosen as a subset of V!). Theorem 6.1 and its generalization in Rem. 6.3 then hold with V replaced by V_h. The same theory as for the continuous problem thus directly gives us existence and uniqueness of a solution u_h to the discrete problem:

$$\text{Find } u_h \in V_h \quad \text{such that} \quad a(u_h, v) = L(v) \quad \text{for all} \quad v \in V_h. \tag{6.16}$$

It remains to obtain an error estimate for the problem on V_h. Let us assume that $a(u, v)$ is symmetric and as such defines a norm (6.10) on V and V_h. Then we have

Lemma 6.4 *The solution u_h to* (6.16) *is the best approximation to the solution u of* (6.3) *for the $\|\cdot\|_a$ norm:*

$$\|u - u_h\|_a = \min_{v \in V_h} \|u - v\|_a. \tag{6.17}$$

Proof. There are two main ideas there. The first idea is that since $V_h \subset V$, we can subtract (6.16) from (6.8) to get the following orthogonality condition

$$a(u - u_h, v) = 0, \qquad \text{for all} \quad v \in V_h. \tag{6.18}$$

Note that this property only depends on the linearity of $u \mapsto a(u, v)$ and not on a being symmetric. Now because $a(u, u)$ is an inner product (so that $a(u, v) \leq \|u\|_a \|v\|_a$) and $v \mapsto a(u, v)$ is linear, we verify that for all $v \in V_h$

$$\|u - u_h\|_a^2 = a(u - u_h, u - u_h) = a(u - u_h, u) = a(u - u_h, u - v) \leq \|u - u_h\|_a \|u - v\|_a.$$

If $\|u - u_h\|_a = 0$, then the lemma is obviously true. If not, then we can divide both sides in the above inequality by $\|u - u_h\|_a$ to get that $\|u - u_h\|_a \leq \|u - v\|_a$ for all $v \in V$. This yields the lemma. □

Remark 6.5 When $a(u, v)$ satisfies the hypotheses given in Rem. 6.3, the above result does not hold. However, it may be replaced by

$$\|u - u_h\|_V \leq \frac{C_1}{\alpha} \min_{v \in V} \|u - v\|_V. \tag{6.19}$$

Exercise 6.2 Prove (6.19).

So, although u_h is no longer the best approximation of u in any norm (because $a(u, v)$ no longer generates a norm), we still observe that up to a multiplicative constant (greater than 1 obviously), u_h is still very close to being an optimal approximation of u (for the norm $\|\cdot\|_V$ now).

Note that the above result also holds when $a(u, v)$ is symmetric. Then u_h is the best approximation of u for the norm $\|\cdot\|_a$, but not for the norm $\|\cdot\|_V$. We have equipped V (and V_h) with two different, albeit equivalent, norms. We recall that two norms $\|\cdot\|_a$ and $\|\cdot\|_V$ are equivalent if there exists a constant C such that

$$C^{-1} \|v\|_V \leq \|v\|_a \leq C \|v\|_V, \qquad \text{for all} \quad v \in V. \tag{6.20}$$

Exercise 6.3 When $a(u, v)$ is symmetric and coercive on V, show that the two norms $\|\cdot\|_a$ and $\|\cdot\|_V$ are equivalent.

Let us summarize our results. We have replaced the variational formulation for $u \in V$ by a discrete variational formulation for $u_h \in V_h$. We have shown that the discrete problem indeed admitted a unique solution (and thus requires to solve a problem of the form $Ax = b$ since the problem is linear and finite dimensional), and that the solution u_h was the best approximation to u when $a(u, v)$ generates a norm, or at least was not too far from the best solution, this time for the norm $\|\cdot\|_V$, when $a(u, v)$ is coercive (whether it is symmetric or not).

In our problem (6.1), $V = H_0^1$. So we have obtained that $\|u - u_h\|_{H^1}$ is close to the best possible approximation of u by functions in H_0^1. Note that the "closeness" is all

relative. For instance, when α is very small, an error bound of the form (6.19) is not necessarily very accurate. When a is symmetric, (6.17) is however always optimal. In any event, we need additional information on the spaces V_h if one wants to go further. We can go further in two directions: see how small $\|u-u_h\|_{H^1}$ is for specific sequences of spaces V_h, but also see how small $u-u_h$ is in possibly other norms one may be interested in, such as $\|u-u_h\|_{L^2}$. We now consider a specific example of Galerkin approximation: the finite element method (although we should warn the reader that many finite element methods are not of the Galerkin form because they are based on discretization spaces V_h that are *not* subspaces of V).

6.3 An example of finite element method

We are now ready to introduce a finite element method. To simplify we assume that Ω is a rectangle. We then divide Ω into a family $\mathcal{T}_h$ of non-overlapping closed triangles. Each triangle $T \in \mathcal{T}_h$ is represented by three vertices and three edges. We impose that two neighboring triangles have one edge (and thus two vertices) in common, or one vertex in common. This precludes the existence of vertices in the interior of an edge. The set of all vertices in the triangulation are called the nodes of $\mathcal{T}_h$. The index $h > 0$ measures the "size" of the triangles. A triangulation $\mathcal{T}_h$ of size h is thus characterized by

$$\bar{\Omega} = \bigcup_{T \in T_h} T, \qquad h_T = \operatorname{diam}(T), \qquad h = \max_{T \in \mathcal{T}_h} h_T. \tag{6.21}$$

As the triangulation gets finer and $h \to 0$, we wish to be able to construct more accurate solutions u_h to u solution of (6.8). Note that finite element methods could also be based on different tilings of Ω, for instance based on quadrilaterals, or higher order polygons. We only consider triangles here.

The reason we introduce the triangulation is that we want to approximate arbitrary functions in $V = H_0^1$ by finite dimensional functions on each triangle $T \in \mathcal{T}_h$. As we have already mentioned, the functions cannot be chosen on each triangle independently of what happens on the other triangles. The reason is that such functions may not match across edges shared by two triangles. These jumps however would violate the fact that our discrete functions need to be in H_0^1. For instance functions that are constant on each triangle $T \in \mathcal{T}_h$ are not in H_0^1 and are thus not admissible. Here we consider the simplest finite element method. The simplest functions after piecewise constant functions are piecewise linear functions. Moreover, we want the functions not to jump across edges of the triangulation. We thus define

$$V_h = \{v \in H_0^1; \quad v \text{ is linear on each } T \in \mathcal{T}_h\}. \tag{6.22}$$

Recall that this imposes that $v = 0$ on $\partial\Omega$. The space V_h is a finite dimensional subspace of H_0^1 (see below for a "proof") so that the theory given in the previous section applies. This gives us a solution u_h to (6.16).

A linear system. We now recast the above problem for u_h as a linear system (of the form $Ax = b$). To do so, we need a basis for V_h. Because the functions in V_h are in H_0^1, then cannot jump across edges. Because they are piecewise linear, they must be

continuous on Ω. So they are characterized (uniquely determined) by the values they take on each node $\mathbf{P}_i \in \Omega$ for $1 \leq i \leq M_h$ of the triangulation $\mathcal{T}_h$ (note that since Ω is an open domain, the notation means that M_h is the number of internal nodes of the triangulation: the nodes $\mathbf{P}_i \in \partial\Omega$ should be labeled with $M_h + 1 \leq i \leq M_h + b_h$, where b_h is the number of nodes $\mathbf{P}_i \in \partial\Omega$). Consider the family of *pyramid functions* $\phi_i(\mathbf{x}) \in V_h$ for $1 \leq i \leq M_h$ defined by

$$\phi_i(\mathbf{P}_j) = \delta_{ij}, \qquad 1 \leq i, j \leq M_h. \tag{6.23}$$

We recall that the Kronecker delta $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Since $\phi_i(\mathbf{x})$ is piecewise affine on $\mathcal{T}_h$, it is uniquely defined by its values at the nodes $\mathbf{P}_j$. Moreover, we easily deduce that the family is linearly independent, whence is a basis for V_h. This proves that V_h is finite dimensional. Note that by assuming that $\phi_i \in V_h$, we have also implicitly implied that $\phi_i(\mathbf{P}_j) = 0$ for $\mathbf{P}_j \in \partial\Omega$. Note also that for $\mathbf{P}_j, \mathbf{P}_k \in \partial\Omega$, the function ϕ_i (for $\mathbf{P}_i \notin \partial\Omega$) vanishes on the whole edge $(\mathbf{P}_j, \mathbf{P}_k)$ so that indeed $\phi_i = 0$ on $\partial\Omega$. We can thus write any function $v \in V_h$ as

$$v(\mathbf{x}) = \sum_{i=1}^{M_h} v_i \phi_i(\mathbf{x}), \qquad v_i = v(\mathbf{P}_i). \tag{6.24}$$

Let us decompose the solution u_h to (6.16) as $u_h(\mathbf{x}) = \sum_{i=1}^{M_h} u_i \phi_i(\mathbf{x})$. The equation (6.16) may then be recast as

$$(f, \phi_i) = a(u_h, \phi_i) = \sum_{j=1}^{M_h} a(\phi_j, \phi_i) u_j.$$

This is nothing but the *linear system*

$$Ax = b, \qquad A_{ij} = a(\phi_j, \phi_i), \quad x_i = u_i, \quad b_i = (f, \phi_i). \tag{6.25}$$

In the specific problem of interest here where $a(u, v)$ is defined by (6.3), we have

$$A_{ij} = \int_\Omega \big(\nabla\phi_j(\mathbf{x}) \cdot \nabla\phi_i(\mathbf{x}) + \sigma(\mathbf{x})\phi_j(\mathbf{x})\phi_i(\mathbf{x})\big) d\mathbf{x}.$$

There is therefore no need to discretize the absorption parameter $\sigma(\mathbf{x})$: the variational formulation takes care of this. Solving for u_h thus becomes a linear algebra problem, which we do not describe further here. Let us just mention that the matrix A is quite sparse because $a(\phi_i, \phi_j) = 0$ unless $\mathbf{P}_i$ and $\mathbf{P}_j$ belong to the same edge in the triangulation. This is a crucial property to obtain efficient numerical inversions of (6.25). Let us also mention that the *ordering* of the points $\mathbf{P}_i$ is important. As a rough rule of thumb, one would like the indices of points $\mathbf{P}_i$ and $\mathbf{P}_j$ belonging to the same edge to be such that $|i - j|$ is as small as possible to render the matrix A as close to a diagonal matrix as possible. The optimal ordering thus very much depends on the triangulation and is by no means a trivial problem.

Approximation theory. So far, we have constructed a family of discrete solutions u_h of (6.16) for various parameters $h > 0$. Let now $\mathcal{T}_h$ be a sequence of triangulations such that $h \to 0$ (recall that h is the maximal diameter of each triangle in the triangulation). It remains to understand how small the error $u - u_h$ is. Moreover, this error will depend on the chosen norm (we have at least three reasonable choices: $\|\cdot\|_a$, $\|\cdot\|_{H^1}$, and $\|\cdot\|_{L^2}$).

Since the problem (6.16) is a Galerkin discretization, we can use (6.17) and (6.19) and thus obtain in the H_0^1 and $a-$ norms that $u - u_h$ is bounded by a constant times $u - v$ for arbitrary $v \in V_h$. Finding the best approximation v for u in a finite dimensional space for a given norm would be optimal. However this is not an easy task in general. We now consider a method, which up to a multiplicative constant (independent of h), indeed provides the best approximation.

Our approximation is based on interpolation theory. Let us assume that $f \in L^2(\Omega)$ so that $u \in H^2(\Omega)$ thanks to (6.14). This implies that u is a continuous function as $2 > d/2 = 1$. We define the interpolation operator $\pi_h : C(\bar{\Omega}) \to V_h$ by

$$(\pi_h v)(\mathbf{x}) = \sum_{i=1}^{M_h} v(\mathbf{P}_i)\phi_i(\mathbf{x}). \tag{6.26}$$

Note that the function v needs to be continuous since we evaluate it at the nodes $\mathbf{P}_i$. An arbitrary function $v \in H_0^1$ need not be well-defined at the points $\mathbf{P}_i$ (although it almost is...)

We now want to show that $\pi_h v - v$ gets smaller as $h \to 0$. This is done by looking at $\pi_h v - v$ on each triangle $T \in \mathcal{T}_h$ and then summing the approximations over all triangles of the triangulation. We thus need to understand the approximation at the level of one triangle. We denote by $\pi_T v$ the affine interpolant of v on a triangle T. For a triangle T with vertices $\mathbf{P}_i$, $1 \le i \le 3$, it is defined as

$$\pi_T v(\mathbf{x}) = \sum_{i=1}^{3} v(\mathbf{P}_i)\phi_i(\mathbf{x}), \quad \mathbf{x} \in T,$$

where the $\phi_i(\mathbf{x})$ are affine functions defined by $\phi_i(\mathbf{P}_j) = \delta_{ij}$, $1 \le i, j \le 3$. We recall that functions in $H^2(T)$ are continuous, so the above interpolant is well-defined for each $v \in H^2(T)$. The various triangles in the triangulation may be very different so we need a method that analyzes $\pi_T v - v$ on arbitrary triangles. This is done in two steps. First we pick a triangle of reference $\hat{T}$ and analyze $\pi_{\hat{T}} v - v$ on it in various norms of interest. Then we define a map that transforms $\hat{T}$ to any triangle of interest T. We then push forward any properties we have found on $\hat{T}$ to the triangle T through this map. The first step goes as follows.

Lemma 6.6 *Let $\hat{T}$ be the triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$ in an orthonormal system of coordinates of $\mathbb{R}^2$. Let $\hat{v}$ be a function in $H^2(\hat{T})$ and $\pi_{\hat{T}}\hat{v}$ its affine interpolant. Then there exists a constant C independent of $\hat{v}$ such that*

$$\begin{aligned} \|\hat{v} - \pi_{\hat{T}}\hat{v}\|_{L^2(\hat{T})} &\le C|\hat{v}|_{H^2(\hat{T})} \\ \|\nabla(\hat{v} - \pi_{\hat{T}}\hat{v})\|_{L^2(\hat{T})} &\le C|\hat{v}|_{H^2(\hat{T})}. \end{aligned} \tag{6.27}$$

We recall that $|\cdot|_{H^2(\hat{T})}$ is the semi-norm *defined in* (6.6).

Proof. The proof is based on judicious use of Taylor expansions. First, we want to replace $\pi_{\hat{T}}$ by a more convenient interpolant $\Pi_{\hat{T}}$. By definition of $\pi_{\hat{T}}$, we deduce that

$$\|\pi_{\hat{T}} v\|_{H^1} \leq C\|v\|_{H^2},$$

since $|v(\mathbf{P}_i)| \leq C\|v\|_{H^2}$ in two space dimensions. Also we verify that for all affine interpolants $\Pi_{\hat{T}}$, we have $\pi_{\hat{T}}\Pi_{\hat{T}} v = \Pi_{\hat{T}} v$ so that

$$\|v - \pi_{\hat{T}} v\|_{H^1} \leq \|v - \Pi_{\hat{T}} v\|_{H^1} + \|\pi_{\hat{T}}(\Pi_{\hat{T}} v - v)\|_{H^1} \leq C\|v - \Pi_{\hat{T}} v\|_{H^1} + C|v|_{H^2}.$$

This is because the second-order derivatives of $\Pi_{\hat{T}} v$ vanish by construction. So we see that it is enough to prove the results stated in the theorem for the interpolant $\Pi_{\hat{T}}$. This is based on the following Taylor expansion (with Lagrange remainder):

$$v(\mathbf{x}) = v(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)\cdot\nabla v(\mathbf{x}_0) + |\mathbf{x} - \mathbf{x}_0|^2 \int_0^1 (\mathbf{x}\cdot\nabla)^2 v((1-t)\mathbf{x}_0 + t\mathbf{x})(1-t)dt. \quad (6.28)$$

Since we will need it later, the same type of expansion yields for the gradient:

$$\nabla v(\mathbf{x}) = \nabla v(\mathbf{x}_0) + |\mathbf{x} - \mathbf{x}_0| \int_0^1 \mathbf{x}\cdot\nabla\nabla v((1-t)\mathbf{x}_0 + t\mathbf{x})dt. \quad (6.29)$$

Both integral terms in the above expressions involve second-order derivatives of $v(\mathbf{x})$. However they may not quite be bounded by $C|v|_{H^2(\hat{T})}$. The reason is that the affine interpolant $v(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)\cdot\nabla v(\mathbf{x}_0)$ and its derivative $\nabla v(\mathbf{x}_0)$ both involve the gradient of v at $\mathbf{x}_0$, and there is no reason for $|\nabla v(\mathbf{x}_0)|$ to be bounded by $C|v|_{H^2(\hat{T})}$. This causes some technical problems. The way around is to realize that $\mathbf{x}_0$ may be seen as a variable that we can vary freely in $\hat{T}$. We therefore introduce the function $\phi(\mathbf{x}) \in C_0^\infty(\hat{T})$ such that $\phi(\mathbf{x}) \geq 0$ and $\int_{\hat{T}} \phi(\mathbf{x})d\mathbf{x} = 1$. It is a classical real-analysis result that such a function exists. Upon integrating (6.28) and (6.29) over $\hat{T}$ in $\mathbf{x}_0$, we obtain that

$$\begin{aligned}
v(\mathbf{x}) &= \Pi_{\hat{T}} v(\mathbf{x}) + \int_{\hat{T}} \phi(\mathbf{x}_0)|\mathbf{x} - \mathbf{x}_0|^2 \int_0^1 (\mathbf{x}\cdot\nabla)^2 v((1-t)\mathbf{x}_0 + t\mathbf{x})(1-t)dtd\mathbf{x}_0 \\
\nabla v(\mathbf{x}) &= \nabla\Pi_{\hat{T}} v(\mathbf{x}) + \int_{\hat{T}} \phi(\mathbf{x}_0)|\mathbf{x} - \mathbf{x}_0| \int_0^1 \mathbf{x}\cdot\nabla\nabla v((1-t)\mathbf{x}_0 + t\mathbf{x})dtd\mathbf{x}_0 \quad (6.30) \\
\Pi_{\hat{T}} v(\mathbf{x}) &= \Big(\int_{\hat{T}} \big(v(\mathbf{x}_0) - \mathbf{x}_0\cdot\nabla v(\mathbf{x}_0)\big)\phi(\mathbf{x}_0)d\mathbf{x}_0\Big) + \Big(\int_{\hat{T}} \phi(\mathbf{x}_0)\nabla v(\mathbf{x}_0)d\mathbf{x}_0\Big)\cdot\mathbf{x}.
\end{aligned}$$

We verify that $\Pi_{\hat{T}} v$ is indeed an affine interpolant for $v \in H^1(\hat{T})$ in the sense (thanks to the normalization of $\phi(\mathbf{x})$) that $\Pi_{\hat{T}} v = v$ when v is a polynomial of degree at most 1 (in x_1 and x_2) and that $\|\Pi_{\hat{T}} v\|_{H^1} \leq C\|v\|_{H^1}$.

It remains to show that the remainders $v(\mathbf{x}) - \Pi_{\hat{T}} v(\mathbf{x})$ and $\nabla(v(\mathbf{x}) - \Pi_{\hat{T}} v(\mathbf{x}))$ are indeed bounded in $L^2(\hat{T})$ by $C|v|_{H^2}$. We realize that both remainders involve bounded functions (such as $|\mathbf{x} - \mathbf{x}_0|$ or $1 - t$) multiplied by second-order derivatives of v. It is therefore sufficient to show a result of the form

$$\int_{\hat{T}} d\mathbf{x}\Big(\int_{\hat{T}} \phi(\mathbf{x}_0)\int_0^1 |f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0))|dtd\mathbf{x}_0\Big)^2 d\mathbf{x}_0 \leq C\|f\|^2_{L^2(\hat{T})}, \quad (6.31)$$

for all $f \in L^2(\hat{T})$. By the Cauchy-Schwarz inequality (stating that $(f,g)^2 \leq \|f\|^2\|g\|^2$, where we use $g = 1$), the left-hand side in the above equation is bounded (because $\hat{T}$ is a bounded domain) by

$$C \int_{\hat{T}} \int_{\hat{T}} \phi^2(\mathbf{x}_0) \int_0^1 f^2((1-t)\mathbf{x}_0 + t\mathbf{x}) dt d\mathbf{x}_0 d\mathbf{x}.$$

We extend $f(\mathbf{x})$ by 0 outside $\hat{T}$. Let us now consider the part $1/2 < t < 1$ in the above integral and perform the change of variables $t\mathbf{x} \leftarrow \mathbf{x}$. This yields

$$\int_{1/2}^1 dt \int_{\hat{T}} d\mathbf{x}_0 \phi^2(\mathbf{x}_0) \int_{t^{-1}\hat{T}} \frac{d\mathbf{x}}{t^2} f^2((1-t)\mathbf{x}_0 + \mathbf{x}) \leq \Big(\int_{1/2}^1 \frac{dt}{t^2} \int_{\hat{T}} d\mathbf{x}_0 \phi^2(\mathbf{x}_0) \Big) \|f\|^2_{L^2(\hat{T})}.$$

This is certainly bounded by $C\|f\|^2_{L^2(\hat{T})}$. Now the part $0 < t < 1/2$ with the change of variables $(1-t)\mathbf{x}_0 \leftarrow \mathbf{x}_0$ similarly gives a contribution of the form

$$\Big(\int_0^{1/2} \frac{dt}{(1-t)^2} \|\phi^2\|_{L^\infty(\hat{T})} |\hat{T}| \Big) \|f\|^2_{L^2(\hat{T})},$$

where $|\hat{T}|$ is the surface of $\hat{T}$. These estimates show that

$$\|v - \Pi_{\hat{T}} v\|_{H^1(\hat{T})} \leq C|v|_{H^2(\hat{T})}. \tag{6.32}$$

This was all we needed to conclude the proof of the lemma. □
Once we have the estimate on $\hat{T}$, we need to obtain it for arbitrary triangles. This is done as follows:

Lemma 6.7 *Let T be a proper (not flat) triangle in the plane $\mathbb{R}^2$. There is an orthonormal system of coordinates such that the vertices of T are the points $(0,0)$, $(0,a)$, and (b,c), where a, b, and c are real numbers such that $a > 0$ and $c > 0$. Let us define $\|A\|_\infty = \max(a, |b|, c)$.*

Let $v \in H^2(T)$ and $\pi_T v$ its affine interpolant. Then there exists a constant C independent of the function v and of the triangle T such that

$$\begin{aligned} \|v - \pi_T v\|_{L^2(T)} &\leq C\|A\|^2_\infty |v|_{H^2(T)} \\ \|\nabla(v - \pi_T v)\|_{L^2(T)} &\leq C\frac{\|A\|^3_\infty}{ac} |v|_{H^2(T)}. \end{aligned} \tag{6.33}$$

Let us now assume that h_T is the diameter of T. Then clearly, $\|A\|_\infty \leq h_T$. Furthermore, let us assume the existence of ρ_T such that

$$\rho_T h_T \leq a \leq h_T, \qquad \rho_T h_T \leq c \leq h_T. \tag{6.34}$$

Then we have the estimates

$$\begin{aligned} \|v - \pi_T v\|_{L^2(T)} &\leq C h_T^2 |v|_{H^2(T)} \\ \|\nabla(v - \pi_T v)\|_{L^2(T)} &\leq C\rho_T^{-2} h_T |v|_{H^2(T)}. \end{aligned} \tag{6.35}$$

We recall that C is independent of v and of the triangle T.

Proof. We want to map the estimate on $\hat{T}$ in Lemma 6.6 to the triangle T. This is achieved as follows. Let us define the linear map $A : \mathbb{R}^2 \to \mathbb{R}^2$ defined in Cartesian coordinates as

$$A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}. \tag{6.36}$$

We verify that $A(\hat{T}) = T$. Let us denote by $\hat{\mathbf{x}}$ coordinates on $\hat{T}$ and by $\mathbf{x}$ coordinates on T. For a function $\hat{v}$ on $\hat{T}$, we can define its push-forward $v = A_*\hat{v}$ on T, which is equivalent to

$$\hat{v}(\hat{\mathbf{x}}) = v(\mathbf{x}) \text{ with } \mathbf{x} = A\hat{\mathbf{x}} \quad \text{or} \quad \hat{v}(\hat{\mathbf{x}}) = v(A\hat{\mathbf{x}}).$$

By the chain rule, we deduce that

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = A_{ik}\frac{\partial v}{\partial x_k},$$

where we use the convention of summation over repeated indices. Differentiating one more time, we deduce that

$$\sum_{|\alpha|=2} |D^\alpha \hat{v}|^2(\hat{\mathbf{x}}) \leq C\|A\|_\infty^4 \sum_{|\alpha|=2} |D^\alpha v|^2(\mathbf{x}), \quad \mathbf{x} = A\hat{\mathbf{x}}. \tag{6.37}$$

Here the constant C is independent of v and $\hat{v}$. Note that the change of measures induced by A is $d\hat{\mathbf{x}} = |\det(A^{-1})|d\mathbf{x}$ so that

$$\int_{\hat{T}} \sum_{|\alpha|=2} |D^\alpha \hat{v}|^2(\hat{\mathbf{x}})d\hat{\mathbf{x}} \leq \|A\|_\infty^4 |\det(A^{-1})| \int_T \sum_{|\alpha|=2} |D^\alpha v|^2(\mathbf{x})d\mathbf{x}.$$

This is equivalent to the statement:

$$|\hat{v}|_{H^2(\hat{T})} \leq C|\det(A)|^{-1/2}\|A\|_\infty^2 |v|_{H^2(T)}. \tag{6.38}$$

This tells us how the right-hand side in (6.27) may be used in estimates on T. Let us now consider the left-hand sides. We first notice that

$$(\pi_{\hat{T}}\hat{v})(\hat{\mathbf{x}}) = (\pi_T v)(\mathbf{x}). \tag{6.39}$$

The reason is that $\hat{v}_i = v_i$ at the nodes and that for any polynomial of order 1 in the variables $\hat{x}_1, \hat{x}_2$, $p(A\hat{\mathbf{x}})$ is also a polynomial of order 1 in the variables x_1 and x_2. This implies that

$$(\hat{v} - \pi_{\hat{T}}\hat{v})(\hat{\mathbf{x}}) = (v - \pi_T v)(\mathbf{x}), \quad A^{-1}\nabla(\hat{v} - \pi_{\hat{T}}\hat{v})(\hat{\mathbf{x}}) = \nabla(v - \pi_T v)(\mathbf{x}); \qquad \mathbf{x} = A\hat{\mathbf{x}}.$$

Using the same change of variables as above, this implies the relations

$$\begin{array}{rcl} \|\hat{v} - \pi_{\hat{T}}\hat{v}\|_{L^2(\hat{T})}^2 & = & |\det A^{-1}| \ \|v - \pi_T v\|_{L^2}^2 \\ \|A^{-1}\|_\infty^2 \|\nabla(\hat{v} - \pi_{\hat{T}}\hat{v})\|_{L^2(\hat{T})}^2 & \geq & C|\det A^{-1}| \ \|\nabla(v - \pi_T v)\|_{L^2}^2. \end{array} \tag{6.40}$$

Here, C is a universal constant, and $\|A^{-1}\|_\infty = (ac)^{-1}\|A\|_\infty$ (i.e., is the maximal element of the 2×2 matrix A^{-1}). The above inequalities combined with (6.38) directly give (6.33). The estimates (6.35) are then an easy corollary. □

Remark 6.8 (Form of the triangles) The first inequality in (6.35) only requires that the diameter of T be smaller than h_T. The second estimate however requires more. It requires that the triangles not be too flat and that all components a, b, and c be of the same order h_T. We may verify that the constraint (6.34) is equivalent to imposing that all the angles of the triangle be greater than a constant independent of h_T and is equivalent to imposing that there be a ball of radius of order h_T inscribed in T. From now on, we impose that $\rho_T \geq \rho_0 > 0$ for all $T \in \mathcal{T}_h$ independent of the parameter h.

This concludes this section on approximation theory. Now that we have local error estimates, we can sum them to obtain global error estimates.

A Global error estimate in $H_0^1(\Omega)$. Let us come back to the general framework of Galerkin approximations. We know from (6.19) that the error of $u-u_h$ in H_0^1 (the space in which estimates are the easiest!) is bounded by a constant times the error $u - v$ in the same norm, where v is the best approximation of u (for this norm). So we obviously have that

$$\|u - u_h\|_{H^1} \leq C\|u - \pi_h u\|_{H^1}. \tag{6.41}$$

Here the constant C depends on the form $a(\cdot,\cdot)$ but not on u nor u_h. The above right-hand side is by definition

$$\Big(\sum_{T\in\mathcal{T}_h}\int_T \big(|u-\pi_T u|^2 + |\nabla(u-\pi_T u)|^2\big)d\mathbf{x}\Big)^{1/2} \leq \Big(\sum_{T\in\mathcal{T}_h} \rho_0^{-4} h_T^2 |u|^2_{H^2(T)}\Big)^{1/2},$$

thanks to (6.35). This however, implies the following result

Theorem 6.9 *Let u be the solution to* (6.8)*, which we assume is in $H^2(\Omega)$. Let u_h be the solution of the discrete problem* (6.16) *where V_h is defined in* (6.22) *based on the triangulation define in* (6.21)*. Then we have the error estimate*

$$\|u - u_h\|_{H^1(\Omega)} \leq C\rho_0^{-2} h |u|_{H^2(\Omega)}. \tag{6.42}$$

Here, C is a constant independent of $\mathcal{T}_h$, u, and u_h. However it may depend on the bilinear form $a(u,v)$. We recall that ρ_0 was defined in Rem. 6.8.

The finite element method is thus first-order for the approximation of u in the H^1 norm. We may show that the order of approximation in $O(h)$ obtained in the theorem is optimal (in the sense that h cannot be replaced by h^α with $\alpha > 1$ in general).

A global error estimate in $L^2(\Omega)$. The norm $\|\cdot\|_a$ is equivalent to the H^1 norm. So the above estimate also holds for the $\|\cdot\|_a$ norm (which is defined only when $a(u,v)$ is symmetric). The third norm we have used is the L^2 norm. It is reasonable to expect a faster convergence rate in the L^2 norm than in the H^1 norm. The reason is this. In the linear approximation framework, the gradients are approximated by piecewise constant functions whereas the functions themselves are approximated by polynomials of order 1 on each triangle. We should therefore have a better accuracy for the functions than for their gradients. This is indeed the case. However the demonstration is not straightforward and requires to use the regularity of the elliptic problem one more time in a curious way.

Theorem 6.10 *Under the hypotheses of Theorem* (6.9) *and the additional hypothesis that* (6.14) *holds, we have the following error estimate*

$$\|u - u_h\|_{L^2(\Omega)} \leq C\rho_0^{-4} h^2 \|u\|_{H^2(\Omega)}. \tag{6.43}$$

Proof. Let w be the solution to the problem

$$a(v, w) = (v, u - u_h), \qquad \text{for all } v \in V. \tag{6.44}$$

Note that since $a(u, v)$ is symmetric, this problem is equivalent to (6.8). However, when $a(u, v)$ is not symmetric, (6.44) is equivalent to (6.8) only by replacing $a(u, v)$ by the *adjoint* bilinear form $a^*(u, v) = a(v, u)$. The equation (6.44) is therefore called a *dual* problem to (6.8), and as a consequence the proof of Theorem 6.10 we now present is referred to as a *duality argument* in the literature.

In any event, the above problem (6.44) admits a unique solution thanks to Theorem 6.1. Moreover our hyp othesis that (6.14) holds implies that

$$\|w\|_{H^2} \leq C\|u - u_h\|_{L^2}. \tag{6.45}$$

Using in that order the choice $v = u - u_h$ in (6.44), the orthogonality condition (6.18), the bound in (6.12) (with $V = H_0^1$), the estimate (6.45), and finally the error estimate (6.42), we get:

$$\begin{aligned} \|u - u_h\|_{L^2}^2 &= a(u - u_h, w) = a(u - u_h, w - \pi_h w) \leq C\|u - u_h\|_{H^1}\|w - \pi_h w\|_{H^1} \\ &\leq C\|u - u_h\|_{H^1}\rho_0^{-2} h\|w\|_2 \leq C\rho_0^{-2} h\|u - u_h\|_{H^1}\|u - u_h\|_{L^2} \\ &\leq C\rho_0^{-4} h^2 \|u\|_{H^2}\|u - u_h\|_{L^2}. \end{aligned}$$

This concludes our result. □

7 Introduction to Monte Carlo methods

7.1 Method of characteristics

Let us first make a small detour by the method of characteristics. Consider a hyperbolic equation of the form

$$\frac{\partial v}{\partial t} - c\frac{\partial v}{\partial x} + \alpha v = \beta, \qquad v(0,x) = P(x), \tag{7.1}$$

for $t > 0$ and $x \in \mathbb{R}$. We want to solve (7.1) by the **method of characteristics**. The main idea is to look for solutions of the form

$$v(t, x(t)), \tag{7.2}$$

where $x(t)$ is a characteristic, solving an ordinary differential equation, and where the rate of change of $\omega(t) = v(t, x(t))$ is prescribed along the characteristic $x(t)$.

We observe that

$$\frac{d}{dt}\Big(v(t,x(t))\Big) = \frac{\partial v}{\partial t} + \frac{dx}{dt}\frac{\partial v}{\partial x},$$

so that the characteristic equations for (7.3) are

$$\frac{dx}{dt} = -c, \qquad \frac{dv}{dt} + \alpha v = \beta. \tag{7.3}$$

The solutions are found to be

$$x(t) = x_0 - ct, \qquad v(t,x(t)) = e^{-\alpha t}v(0,x_0) + \frac{\beta}{\alpha}(1 - e^{-\alpha t}). \tag{7.4}$$

For each (t,x), we can find a unique $x_0(t,x) = x + ct$ so that

$$v(t,x) = e^{-\alpha t}P(x+ct) + \frac{\beta}{\alpha}(1 - e^{-\alpha t}). \tag{7.5}$$

Many first-order linear and non-linear equations may be solved by the method of characteristics. The main idea, as we have seen, is to replace a partial differential equation (PDE) by a system of ordinary differential equations (ODEs). The latter may then be discretized if necessary and solved numerically. Since ODEs are solved, we avoid the difficulties in PDE-based discretizations caused by the presence of high frequencies that are not well captured by the schemes. See the section on finite difference discretizations of hyperbolic equations.

There are many other difficulties in the numerical implementation of the method of characteristics. For instance, we need to solve the characteristic backwards from $x(t)$ to $x(0)$ and evaluate the initial condition at $x(0)$, which may not be a grid point and thus may require that we use an interpolation technique. Yet the method is very powerful as it is much easier to solve ODEs than PDEs. In the literature, PDE-based, grid-based, methods are referred to as **Eulerian**, whereas methods acting in the reference frame of the particles following their characteristic (which requires solving ODEs) are referred to as **Lagrangian** methods. The combination of grid-based and particle-based methods gives rise to the so-called *semi-Lagrangian* method.

7.2 Probabilistic representation

One of the main drawbacks of the method of characteristics is that it applies (only) to scalar first-order equations. In one dimension of space, we can apply it to the wave equation

$$0 = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = (\frac{\partial}{\partial t} - \frac{\partial}{\partial x})(\frac{\partial}{\partial t} + \frac{\partial}{\partial x})u,$$

because of the above decomposition. However, the method does not apply to multi-dimensional wave equations nor to elliptic or parabolic equations such as the Laplace and heat equations. The reason is that information does not propagate along characteristics (curves).

For some equations, a quasi-method of characteristics may still be developed. However, instead of being based on us following one characteristics, it is based on us following an infinite number of characteristics generated from an appropriate probability measure and *averaging* over them. More precisely, instead of looking for solutions of the form $w(t, x(t))$, we look for solutions of the form

$$u(t, x) = \mathbb{E}\{u_0(X(t))|X(0) = x\}, \tag{7.6}$$

where $X(t) = X(t, \omega)$ is a trajectory starting at x for realizations ω in a state space Ω, and $\mathbb{E}$ is ensemble averaging with respect to a measure $\mathbb{P}$ defined on Ω. It is not purpose of these notes to dwell on (7.6). Let us give the most classical of examples. Let $W(t)$ be a standard one-dimensional Brownian motion with $W(0) = 0$. This is an object defined on a space Ω (the space of continuous trajectories on $[0, \infty)$) with an appropriate measure $\mathbb{P}$. All trajectories are therefore continuous (at least almost surely) but are also almost surely always non-differentiable. The "characteristics" have thus become rather un-smooth objects. In any event, we find that

$$u(t, x) = \mathbb{E}\{u_0(x + W(2t))\}, \tag{7.7}$$

solves the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad u(0, x) = u_0(x). \tag{7.8}$$

This may be shown by e.g. Itô calculus. Very briefly, let $g(W(2t)) = u_0(x + W(2t))$. Then, we find that

$$dg = \frac{\partial g}{\partial x} dW_{2t} + \frac{1}{2}\frac{\partial^2 g}{\partial x^2} dW_{2t} dW_{2t} = \frac{\partial g}{\partial x} dW_{2t} + \frac{\partial^2 g}{\partial x^2} dt,$$

since $dW_{2t} dW_{2t} = d(2t)$. Using $\mathbb{E}\{W_{2t}\} = 0$, we obtain after ensemble averaging that (7.7) solves (7.8). We thus see that $u(t, x)$ is an infinite superposition ($\mathbb{E}$ is an integration) of pieces of information carried along the trajectories. We also see how this representation gives rise to the Monte Carlo method. We cannot simulate an infinite number of trajectories. However, we can certainly simulate a finite, large, number of them and perform the averaging. This is the basis for the Monte Carlo method. Note that in (7.7), simulating the characteristic is easy as we know the law of $W(t)$ (a centered Gaussian variable with variance equal to t). In more general situations, where the diffusion coefficient depends on position, then $W(2t)$ is replaced by some process $X(t; x)$ which depends in a more complicated way on the location of the origin x. Solving for the trajectories becomes more difficult.

7.3 Random Walk and Finite Differences

Brownian motion has a relatively simple discrete version, namely the random walk. Consider a process S_k such that $S_0 = 0$ and S_{k+1} is equal to $S_k + h$ with probability $\frac{1}{2}$ and equal to $S_k - h$ also with probability $\frac{1}{2}$.

We may decompose this random variable as follows. Let us assume that we have N (time) steps and define N independent random variables x_k, $1 \leq k \leq N$, such that $x_k = \pm 1$ with probability $\frac{1}{2}$. Then,

$$S_n = h \sum_{k=1}^{n} x_k,$$

is our random walk.

Let us relate all this to Bernoulli random variables. Define y_k for $1 \leq k \leq N$, N independent random variables equal to 0 or 1 with probability $\frac{1}{2}$ and let $\Sigma_n = \sum_{k=1}^{n} y_k$. Each sequence of n of 0's and 1's for the y_k has probability 2^{-n}. The number of sequences such that $\Sigma_n = m$ is $\binom{n}{m}$ by definition (number of possibilities of choosing m balls among $n \geq m$ balls). As a consequence, using $x_k = (2y_k - 1)$, we see that

$$p_m = \mathbb{P}(S_n = hm) = \frac{1}{2^n} \binom{n}{\frac{n+m}{2}}. \tag{7.9}$$

Exercise 7.1 Check this.

Note that S_n is even when n is even and odd when n is odd so that $\frac{n+m}{2}$ is always an integer. We thus have an explicit expression for S_n. Moreover, we easily find that

$$\mathbb{E}\{S_n\} = 0, \qquad \mathrm{Var} S_n = \mathbb{E}\{S_n^2\} = nh^2. \tag{7.10}$$

The central limit theorem shows that S_n converges to $W(2t)$ if $n \to \infty$ and $h \to 0$ such that $nh^2 = 2t$. Moreover, $S_{\alpha n}$ also converges to $W(2\alpha t)$ in an appropriate sense so that there is solid mathematical background to obtain that S_n is indeed a bona fide discretization of Brownian motion. This is Donsker's invariance principle.

How does one use all of this to solve a discretized version of the heat equation? Following (7.7), we can define

$$U_j^n = \mathbb{E}\{U^0_{j+h^{-1}S_n}\}, \tag{7.11}$$

where U_j^0 is an approximation of $u_0(jh)$ as before, and where U_j^n is an approximation of $u(n\Delta t, jh)$. What is it that we have defined? For this, we need to use the idea of conditional expectations, which roughly says that averaging over two random variables may be written as an averaging over the first random variable (this is thus still a random variable) and then average over the second random variable. In our context:

$$\begin{aligned} U_j^n &= \mathbb{E}\{U^0_{j+x_1+\ldots+x_n}\} = \mathbb{E}\{\mathbb{E}\{U^0_{j+x_1+\ldots+x_n}|x_n\}\} \\ &= \frac{1}{2}\mathbb{E}\{U^0_{j+1+x_1+\ldots+x_{n-1}}\} + \frac{1}{2}\mathbb{E}\{U^0_{j-1+x_1+\ldots+x_{n-1}}\} = \frac{1}{2}\Big(U_{j+1}^{n-1} + U_{j-1}^{n-1}\Big). \end{aligned}$$

In other words, U_j^n solves the finite difference equation with $\lambda = \frac{1}{2}$ (and ONLY this value of λ the way the procedure is constructed). Recall that

$$\lambda = \frac{D\Delta t}{h^2} = \frac{1}{2}, \tag{7.12}$$

where D is the diffusion coefficient. In other words, the procedure (7.11) gives us an approximation of

$$\frac{\partial u}{\partial t} - D\frac{\partial^2 u}{\partial x^2} = 0, \qquad u(0,x) = u_0,$$

when Δt and h are chosen so that (7.12) holds.

7.4 Monte Carlo method

Solving for U_j^n thus requires that we construct S_n (by flipping n ± 1 coins) and evaluate $X := U^0_{j+h^{-1}S_n}$. If we do it once, how accurate are we? The answer is not much. Let us look at the random variable X. We know that $\mathbb{E}\{X\} = U_j^n$ so that X is an unbiased estimator of U_j^n. However, its variance is often quite large. Using (7.9), we find that $\mathbb{P}(X = U^0_{j+m}) = p_m$. As a consequence, we find that:

$$\sigma^2 = \mathrm{Var}X = \sum_{k=1}^{n} p_k \Big(U^0_{j-k} - \sum_{l=1}^{n} p_l U^0_{j-l} \Big)^2. \tag{7.13}$$

Assume that $U_j^0 = U$ independent of j. Then $\sigma^2 = 0$ and X gives the right answer all the time. However, when U^0 is highly oscillatory, say $(-1)^l$, then $\sum p_l U^0_{j-l}$ will be close to 0 by cancellations. This shows that σ^2 will be close to $\sum_k p_k (U^0_{j-k})^2$, which is comparable, or even much larger than $\mathbb{E}\{X\}$. So X has a very large variance and is therefore not a very good estimator of U_j^n.

The (only) way to make the estimator better is to use the law of large numbers by repeating the experiment M times and averaging over the M results. More precisely, let X_k for $1 \le k \le M$ be M independent random variables with the same law as X. Define

$$S_M = \frac{1}{M} \sum_{k=1}^{M} X_k. \tag{7.14}$$

Then we verify that $\mathbb{E}\{S_M\} = U_j^n$ as before. The variance of S_M is however significantly smaller:

$$\mathrm{Var}S_M = \mathbb{E}\{ \Big(\frac{1}{M} \sum_{k=1}^{M} (X_k - U_j^n) \Big)^2 \} = \frac{1}{M^2} \mathbb{E}\{ \sum_{k=1}^{M} (X_k - U_j^n)^2 \} = \frac{\mathrm{Var}X}{M},$$

since the variables are independent. As a consequence, S_M is an unbiased estimator of U_n^j with a standard deviation of order $O(M^{-1/2})$. This is the speed of convergence of Monte Carlo. In order to obtain an accuracy of ε on average, we need to calculate an order of ε^{-2} realizations.

How do the calculations Monte Carlo versus deterministic compare to calculate U_j^n (at a final time $n\Delta t$ and a *given* position j)? Let us assume that we solve a d-dimensional

heat equation, which involves a d-dimensional random walk, which is nothing but d one-dimensional random walks in each direction. The cost of each trajectory is of order $n \approx (\Delta t)^{-1}$ since we need to flip n coins. To obtain an $O(h^2) = O(\Delta t)$ accuracy, we thus need $M = (\Delta t)^{-2}$ so that the final cost is $O((\Delta t)^{-3})$.

Deterministic methods require that we solve n time steps using a grid of mesh size $O(h)$ in each direction, which means $O(h^{-d})$ discretization points. The total cost of an explicit method is therefore $O((\Delta t h^d)^{-1}) = O((\Delta t)^{-1-\frac{d}{2}})$. We thus see that Monte Carlo is more efficient than finite differences (or equally efficient) when

$$\frac{1}{\Delta t^3} \leq \frac{1}{\Delta t^{1+\frac{d}{2}}}, \qquad \text{i.e.,} \quad d \geq 4. \tag{7.15}$$

For sufficiently large dimensions, Monte Carlo methods do not suffer the *curse of dimensionality* characterizing deterministic methods and thus become more efficient. Monte Carlo methods are also much more versatile as they are based on solving "random ODEs" (stochastic ODEs), with which it is easier to account for e.g. complicated geometries. Finite differences however provide the solution everywhere unlike Monte Carlo methods, which need to use different trajectories to evaluate solutions at different points.

Moreover, the Monte Carlo method is much easier to parallelize. For the finite difference model, it is difficult to use parallel architecture (though this a well-studied problem) and obtain algorithms running time is inversely proportional to the number of available processors. In contrast, Monte Carlo methods are easily parallelized since the M random realizations of X are done independently. So with a (very large) number of processors equal to M, the running time of the algorithm is of order Δt^{-1}, which corresponds to the running time of the finite difference algorithm on a grid with a small finite number of grid points. Even with a large number of processors, it is extremely difficult to attain this level of running speed using the deterministic finite difference algorithm.

Let us conclude by the following remark. The convergence of the Monte Carlo algorithm in $M^{-\frac{1}{2}}$ is rather slow. Many algorithms have been developed to accelerate convergence. These algorithms are called variance-reduction algorithms and are based on estimating the solution U_j^n by using non-physical random processes with smaller variance than the classical Monte Carlo method described above.

References

[1] S. Larsson and V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer, New York, 2003.

[2] L. N. Trefethen, *Spectral Methods in Matlab*, SIAM, Philadelphia, PA, 2000.