Lecture Notes on
COMPUTATIONAL PHYSICS

Emmanuel Dormy

Ecole Normale Supérieure
Physics Department

Chapter 1

Where the troubles begin...

1.1 Introduction

Most of the laws of macroscopic physics (and at any rate, all laws considered in these notes) are expressed in terms of partial differential equations. These relate physical quantities to their derivatives in space and in time.

When possible, physicists will most often seek analytical solutions to these equations. However in most cases such solutions do not exist... One then has to resort to trying to get apporached solutions using a computer.

The laws of physics are usually described by partial differential equations (PDEs). This includes many types of applications: physics of materials, fluid mechanics, electromagnetism ... granular media are an exception as they obey different rules. One can sometimes find analytical solutions of PDEs. However most of the time there are no analytical solutions. When this is the case, one must resort to seek approximate solutions on machines which, unfortunately, do not like "infinity". This raises two issues:

- First, functions correspond to an infinite number of degrees of freedom (or unknown). In order to derive a problem involving only a finite number of unknown, it is necessary to "discretise" the problem.
- Even once discretised, the equations involve real numbers... In order to represent these numbers using only a finite amount of information ("bits" in a computer) one must introduce rounding errors.

Let us consider in somewhat formal manner the transition from a continuous to a discrete problem. We can formally express our original continuous problem in the form

$$\begin{cases} L(\boldsymbol{u}) = \boldsymbol{g} & \text{in} \quad \Omega\,, \\ B(\boldsymbol{u}) = \boldsymbol{\gamma} & \text{on} \quad \partial\Omega\,, \end{cases} \tag{1.1}$$

where $\boldsymbol{u}$ stands for the unknown function. L and B representing differential operators respectively corresponding to the PDE in the domain of interest and the boundary conditions (which may also involve partial derivatives).

The discrete problem can then be expressed in the form

$$\begin{cases} L_h(\boldsymbol{u}_h) = \boldsymbol{g} & \text{in} \quad \Omega\,, \\ B_h(\boldsymbol{u}_h) = \boldsymbol{\gamma} & \text{on} \quad \partial\Omega\,, \end{cases} \tag{1.2}$$

where the subscript h identifies the typical discretisation step.
As we will ponder on the many possible ways to construct a discrete problem, a few questions should guide our thinking:

- Does L_h offer a good approximation to L ?
  In more mathematical terms, does $L_h \underset{h\to 0}{\longrightarrow} L$?
- Does $\boldsymbol{u}_h$ offer a good approximation to $\boldsymbol{u}$?
  Does $\boldsymbol{u}_h \underset{h\to 0}{\longrightarrow} \boldsymbol{u}$?

The first concern refers to a notion we will call *consistancy*, while the second will be named *convergence*.

The truncation error can then be defined as

$$R_h(\boldsymbol{u}) = L_h(\boldsymbol{u}) - L(\boldsymbol{u}) = L_h(\boldsymbol{u}) - \boldsymbol{g}\,. \tag{1.3}$$

Another way to express consistancy is then $R_h(\boldsymbol{u}) \underset{h\to 0}{\longrightarrow} 0\,.$

The distinction between consistancy and convergence may seem a bit technical at this stage. The reader may even wonder whether it was really necessary to introduce two notions. We will soon find out that the distinction between these notions is quite essential and that they should not be confused.

1.2 A first example

We will consider a problem of thermal convection. This can be studied experimentally and we will try to build a numerical model.

Considering the Fourier law

$$\boldsymbol{q} = -k\boldsymbol{\nabla}T \quad \text{where } k \text{ is the thermal conductivity,} \tag{1.4}$$

the temporal evolution of the energy is

$$\rho\frac{\partial e}{\partial t} = -\boldsymbol{\nabla}\cdot\boldsymbol{q} \quad \text{where } e = cT \text{ and } c \text{ is the calorific capacity.} \tag{1.5}$$

Assuming k to be constant in space and introducing $\kappa = k/\rho c$, one gets

$$\frac{\partial T}{\partial t} = \kappa\,\Delta T\,. \tag{1.6}$$

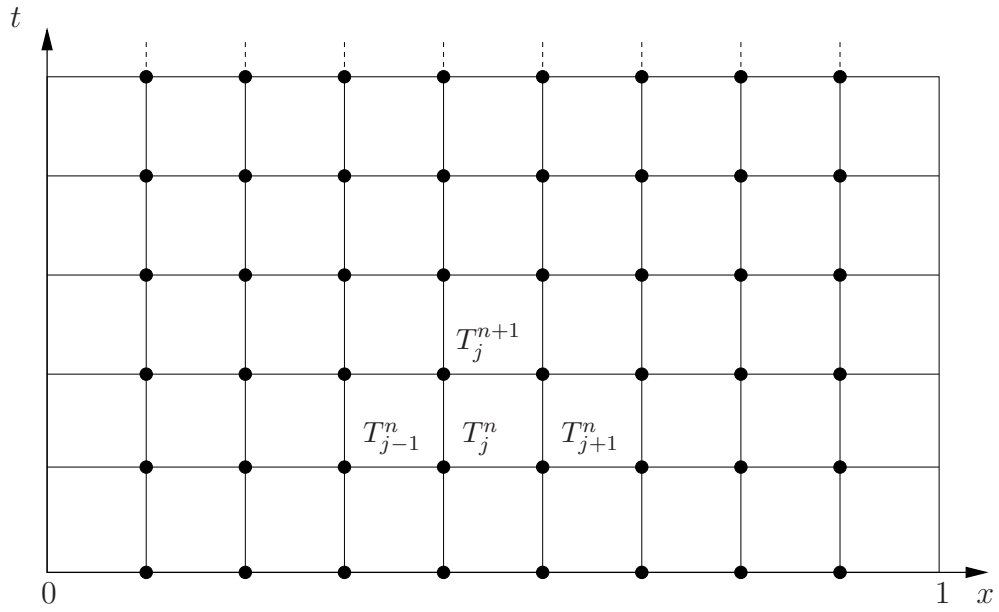

Figure 1.1: Regular grid in space and time.

We can further simplify the problem by modeling a one spatial dimension problem (the vertical dimension). Consider first that the temperature is kept constant at both sides. We can then write for $x \in [0, 1]$

$$\begin{cases} \dfrac{\partial T}{\partial t} = \kappa \dfrac{\partial^2 T}{\partial x^2}\,, \\ T = 0\,, \qquad \text{at } x = 0 \text{ and } x = 1,\, \forall t\,, \\ \\ T = \sin(2\pi\, x)\,, \qquad \text{at } t = 0\,. \end{cases} \tag{1.7}$$

If one discretises time using a step Δt and space using a step Δx, each unknown can be identified using its indices j in space and n in time. If N is the number of unknowns in space (values at $x = 0$ and $x = 1$ are known), we have $\Delta x = 1/(N + 1)$ and $x = j\,\Delta x$, $t = n\,\Delta t$.

We can intuitively derive the following scheme

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \kappa\, \frac{\frac{T_{j+1}^n - T_j^n}{\Delta x} - \frac{T_j^n - T_{j-1}^n}{\Delta x}}{\Delta x}\,, \tag{1.8}$$

which we can rewrite

$$T_j^{n+1} = T_j^n + c\,(T_{j-1}^n - 2\,T_j^n + T_{j+1}^n)\,, \qquad \text{with} \quad c \equiv \frac{\Delta t\,\kappa}{\Delta x^2}\,. \tag{1.9}$$

This stencil can be derived more formally by writting a Taylor expansion in space

$$\begin{aligned} T^n_{j+\alpha} &= T^n_j + \alpha\,\Delta x \left(\frac{\partial T}{\partial x}\right)^n_j + \alpha^2\,\frac{\Delta x^2}{2}\left(\frac{\partial^2 T}{\partial x^2}\right)^n_j + \alpha^3\,\frac{\Delta x^3}{3!}\left(\frac{\partial^3 T}{\partial x^3}\right)^n_j \\ &\quad +\alpha^4\,\frac{\Delta x^4}{4!}\left(\frac{\partial^4 T}{\partial x^4}\right)^n_j + \alpha^5\,\frac{\Delta x^5}{5!}\left(\frac{\partial^5 T}{\partial x^5}\right)^n_j + \mathcal{O}(\Delta x^6)\,, \end{aligned} \tag{1.10}$$

and a similar expansion in time

$$\begin{aligned} T^{n+\alpha}_j &= T^n_j + \alpha\,\Delta t \left(\frac{\partial T}{\partial t}\right)^n_j + \alpha^2\,\frac{\Delta t^2}{2}\left(\frac{\partial^2 T}{\partial t^2}\right)^n_j + \alpha^3\,\frac{\Delta t^3}{3!}\left(\frac{\partial^3 T}{\partial t^3}\right)^n_j \\ &\quad +\alpha^4\,\frac{\Delta t^4}{4!}\left(\frac{\partial^4 T}{\partial t^4}\right)^n_j + \alpha^5\,\frac{\Delta t^5}{5!}\left(\frac{\partial^5 T}{\partial t^5}\right)^n_j + \mathcal{O}(\Delta t^6)\,, \end{aligned} \tag{1.11}$$

(with $\alpha \in I\!N$). Adding the expressions corresponding to (1.10) with $\alpha = 1$ and $\alpha = -1$, all the odd numbers cancel out and we are left with

$$T_{j-1} + T_{j+1} = 2T_j + \Delta x^2\,\left.\frac{\partial^2 T}{\partial x^2}\right|^n_j + \frac{\Delta x^4}{12}\,\left.\frac{\partial^4 T}{\partial x^4}\right|^n_j + \mathcal{O}(\Delta x^6)\,, \tag{1.12}$$

thus the approximation

$$\left.\frac{\partial^2 T}{\partial x^2}\right|^n_j = \frac{T^n_{j-1} - 2T^n_j + T_{j+1^n}}{\Delta x^2} - \frac{\Delta x^2}{12}\,\left.\frac{\partial^4 T}{\partial x^4}\right|^n_j + \mathcal{O}(\Delta x^4)\,. \tag{1.13}$$

We decduce from this short calculation that the stencil we proposed in (1.9) is accurate to second order in space.

The same reasonning using (1.11) with $\alpha = 1$, yields

$$\left.\frac{\partial T}{\partial t}\right|^n_j = \frac{T^{n+1}_j - T^n_j}{\Delta t} - \frac{\Delta t}{2}\,\left.\frac{\partial^2 T}{\partial t^2}\right|^n_j - \mathcal{O}(\Delta t^2)\,. \tag{1.14}$$

Our stencil is thus only first order accurate in time (i.e. the truncation error term vanishes to 0 as Δt).

We can now identify our expressions with the general formalism we just introduced, then

$$L(T) = \frac{\partial T}{\partial t} - \kappa\frac{\partial^2 T}{\partial x^2} = 0\,, \tag{1.15}$$

and we can then write

$$\begin{aligned} L(T) &= \frac{T(x_j, t_{n+1}) - T(x_j, t_n)}{\Delta t} - \kappa\frac{T(x_{j-1}, t_n) - 2T(x_j, t_n) + T(x_{j+1}, t_n)}{\Delta x^2} \\ &\quad -\frac{\Delta t}{2}\,\left.\frac{\partial^2 T}{\partial t^2}\right|^n_j + \kappa\frac{\Delta x^2}{12}\,\left.\frac{\partial^4 T}{\partial x^4}\right|^n_j + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4)\,, \end{aligned} \tag{1.16}$$

and thus

$$R_h(T) = \frac{\Delta t}{2}\left.\frac{\partial^2 T}{\partial t^2}\right|_j^n - \kappa\frac{\Delta x^2}{12}\left.\frac{\partial^4 T}{\partial x^4}\right|_j^n + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4)\,. \tag{1.17}$$

A convergence analysis on the computer however reveals that the error does not always vanish...

1.3 Von Neumann stability analysis

We ensured that the discrete equation was "close" to the initial problem, the difference being defined as the truncation error. However, during the numerical resolution, we will rely on this approximation of the continuous problem by the discrete stencil a high number of times. Small errors performed at each time step could therefore accumulate and the numerical solution could then deviate from the solution of the original problem.

An efficient approach to investigate the numerical stability of a stencil is to perform the Von Neumann stability analysis. Let us consider an infinite, or periodic, domain, with a regular grid $x_j = j\Delta x$ and a harmonic pertubation at a given time of the form $f = \mathrm{e}^{ikx}$.

The time evolution of this perturbation can be established by considering the amplification factor, which we define as

$$\xi = \frac{f^{n+1}}{f^n}\,. \tag{1.18}$$

This complex factor multiplies the considered perturbation at each time step. By injecting this expressions in (1.9), we get

$$\xi = 1 + c\,(\mathrm{e}^{ik\Delta x} - 2 + \mathrm{e}^{-ik\Delta x})\,, \tag{1.19}$$

with $\mathrm{e}^{ik\Delta x} + \mathrm{e}^{-ik\Delta x} = 2\cos(k\Delta x)\,,$

$$\xi = 1 - 2\,c\,\left[1 - \cos k\Delta x\right]\,, \tag{1.20}$$

and $2\,\sin^2 x = 1 - \cos 2x$, thus we get

$$\xi = 1 - 4\,c\,\sin^2\left(\frac{k\Delta x}{2}\right)\,. \tag{1.21}$$

For the perturbation not to be amplified in time, it is necessary to require that $|\xi| \leq 1$.

In our case $\xi \in I\!R$ and we always have $\xi < 1$. To ensure that $-1 \leq \xi$, one can consider the wake number k which maximises the sinus, one then gets

$$-1 \;\leq\; 1 - 4c\,, \tag{1.22}$$

$$c \;\equiv\; \kappa\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}\,. \tag{1.23}$$

Even if Δt and Δx are small, if they do not match this criterion, the scheme, eventhough consistant will not offer a good approximation to our original continuous problem, as it will amplify perturbations.

1.4 The Lax theorem

Can we be sure to have identified all the possible difficulties?
Are the consistancy and the stability the two only relevant properties?

Lax theorem :

For a well-posed linear initial value problem, wellposed in the sens of Hadamard[1] *consistancy* and *stability* of the numerical scheme imply the *convergence* of the solution of the discrete problem to that of the original problem.

[1] A problem is said to be well-posed in the sens of Hadamard if there exists one, and only one, solution which changes smoothly when the parameters are varied.

Chapter 2

Taylor expansion and Finite Differences

2.1 Stencils and orders

Relying on the Taylor expansion approach introduced in the previous chapter [see equation (1.10)], one can straightforwardly derive many stencils (or schemes) to approximate, say, the first derivative

$$\partial_x u|_j = \frac{u_{j+1}-u_{j-1}}{2\Delta x} - \frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_j + \dots, \tag{2.1}$$

$$\partial_x u|_j = \frac{u_j-u_{j-1}}{\Delta x} + \frac{\Delta x}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_j - \frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_j + \dots, \tag{2.2}$$

$$\partial_x u|_j = \frac{u_{j+1}-u_j}{\Delta x} - \frac{\Delta x}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_j - \frac{\Delta x^2}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_j + \dots. \tag{2.3}$$

Let us introduce the following notations to identify the three above schemes

$$\delta_o u\,|_j = \frac{u_{j+1}-u_{j-1}}{2\Delta x}, \qquad \delta_- u\,|_j = \frac{u_j-u_{j-1}}{\Delta x}, \text{and } \delta_+ u\,|_j = \frac{u_{j+1}-u_j}{\Delta x}.$$

The very fact that we are given the choice between three competing schemes to represent one given derivative is in it self a bit alarming. Surely the choice is not obvious and will not be free of consequences...

Having established the above formula we are naturally pushed by the irresistible temptation to involve more neighbouring points. One can for example use three points and write the non-symmetric second order stencils

$$\partial_x u|_j = \frac{-3u_j+4u_{j+1}-u_{j+2}}{2\Delta x} + \frac{\Delta x^2}{3}\left(\frac{\partial^3 u}{\partial x^3}\right) + \dots, \tag{2.4}$$

$$\partial_x u|_j = \frac{3u_j-4u_{j-1}+u_{j-2}}{2\Delta x} + \frac{\Delta x^2}{3}\left(\frac{\partial^3 u}{\partial x^3}\right) + \dots, \tag{2.5}$$

or widen the stencil and write the four points fourth order formula

$$\partial_x u|_j = \frac{-u_{j+2} + 8u_{j+1} - 8u_{j-1} + u_{j+2}}{12\Delta x} + \frac{\Delta x^4}{3}\left(\frac{\partial^5 u}{\partial x^5}\right) + \dots . \tag{2.6}$$

Following a similar approach, schemes of various order can be written for the second order derivative

$$\partial_{xx} u|_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{\Delta x^2} - \frac{\Delta x^2}{12}\left(\frac{\partial^4 u}{\partial x^4}\right) + \dots , \tag{2.7}$$

$$\partial_{xx} u|_j = \frac{-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}}{12\Delta x^2} + \frac{\Delta x^4}{90}\left(\frac{\partial^6 u}{\partial x^6}\right) + \dots . \tag{2.8}$$

These are only a few (probably the most classical) of the possible stencils. The methodology introduced here makes it straightforward to construct further expressions, either involving a different choice of points or representing higher derivatives.

2.2 Lagrange polynomials

An alternative approach to derive finite difference formula is to seek for a polynomial expression of degree N through $N+1$ neighbouring points. As an example, assume that we know $u(0) = u_0$, $u(\Delta x) = u_1$, $u(2\Delta x) = u_2$, and seek for a polynomial of the form $u(x) = a + b\,x + c\,x^2$, through these points.

We can easily get $$b = \frac{-3u_0 + 4u_1 - u_2}{2\Delta x}\,, \qquad 2\,c = \frac{u_0 - 2u_1 + u_2}{\Delta x^2}\,.$$

Computing the exact expressions for the derivative of this polynomial at point $x = 0$ yields the second order expression corresponding to (2.4). A first order expression for the second derivative can also easily be derived from c.

A general way to proceed is to use the Lagrange polynomial. Assuming the points

$$(x_{j-1}, f_{j-1})\,, \qquad (x_j, f_j)\,, \qquad (x_{j+1}, f_{j+1})\,,$$

are known, the polynomial takes the form

$$F(x) = \frac{(x - x_j)(x - x_{j+1})}{(x_{j-1} - x_j)(x_{j-1} - x_{j+1})}\, f_{j-1} + \frac{(x - x_{j-1})(x - x_{j+1})}{(x_j - x_{j-1})(x_j - x_{j+1})}\, f_j + \frac{(x - x_{j-1})(x - x_j)}{(x_{j+1} - x_{j-1})(x_{j+1} - x_{j-1})}\, f_{j+1}\,. \tag{2.9}$$

It corresponds to the sum of three parabola, respectively passing through the points $((x_{j-1}, f_{j-1}), (x_j, 0), (x_{j+1}, 0))$, $((x_{j-1}, 0), (x_j, f_j), (x_{j+1}, 0))$ and $((x_{j-1}, 0), (x_j, 0), (x_{j+1}, f_{j+1}))$.

This provides a systematic approach to derive the finite difference coefficients (which can easily be extended to stretched grids). It is however important to note that this is **not** a reconstruction of the function (as illustrated in figure 2.1).

2.3 Dual grid and staggered meshes

The notion of dual grid is essential in computational physics. It is simple, but occures very often and sometimes unexpectedly. It is directly related to the fact that three different expressions (δ_0, δ_+, and δ_-) were obtained for the first derivative when involving only neighbouring points. The best way to understand it is to appreciate that a minimal definition of the first derivative would relate to the slope of the line (1st order polynomial) passing through two points. It is easy to see that this expression is second order only at the center of the interval defined by these two points. As a result the first derivative is naturally defined between the grid points at which the function is known.

To illustrate this, let us now assume that we now consider a problem in which the thermal conductivity coefficient is varying in space

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(\kappa(x)\frac{\partial T}{\partial x}\right)\,. \tag{2.10}$$

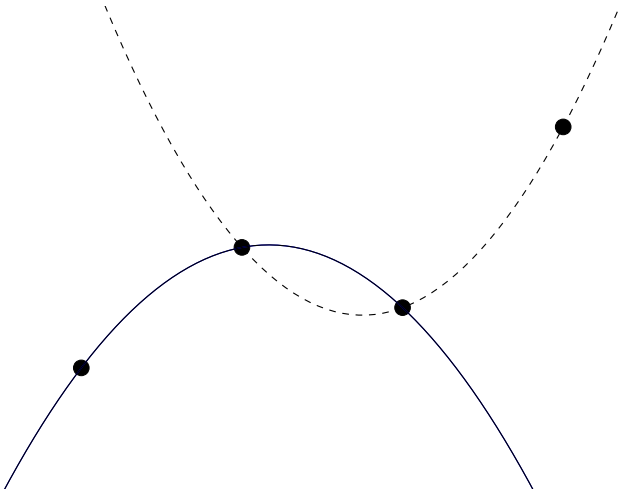

Figure 2.1: Lagrange polynomials offer an easy way to derive finite difference formula they however clearly do not correspond to a reconstruction of the function between the discretisation points. The figure highlights that it does not provide a unique reconstruction between two neighbouring points.

It is then tempting to develop this expression as

$$\frac{\partial T}{\partial t} = \kappa(x)\frac{\partial^2 T}{\partial x^2} + \frac{\partial \kappa(x)}{\partial x}\frac{\partial T}{\partial x}\,. \tag{2.11}$$

The first term is familiar. The second term, however involves first derivatives. This implies a selection between δ_0, δ_+, δ_- which is far from obvious. Such analytical development is to be avoided, and it is much more appropriate to discretise (2.10) rather than (2.11).

In order to achieve this, it is useful to note that combinations of δ_+ and δ_- provide

$$\delta_-\delta_+ = \delta_+\delta_- = \frac{u_{j-1} - 2u_j + u_{j+1}}{\Delta x^2}\,, \tag{2.12}$$

i.e. the second order formula (2.7) for the second derivative that we used so far. It is interesting that these combinations of two first order stencils result in a second order approximation. The combination of the second order formula with itself $\delta_o\delta_o$ would instead result in a sparse stencil (by sparse we mean that it "skips" the direct neighbours)

$$\delta_0\delta_0 = \frac{u_{j-2} - 2u_j + u_{j+2}}{(2\,\Delta x)^2}\,. \tag{2.13}$$

This is to be avoided as it introduces decoupled grids.

If one uses the above centered scheme and define each quantity (T and q) at all points, one can easily see that the scheme results in two decoupled grids (see figure 2.2): one involving T at odd indices and q at even indices, and the other one involving the reverse. This is not only a waist of computational power (the computation could be performed with

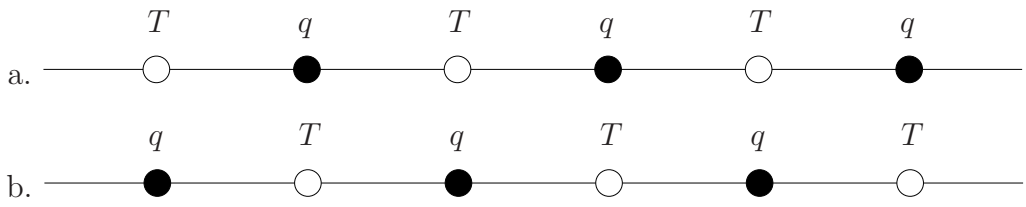

Figure 2.2: Two decoupled grids when using the centered scheme in one dimension of space.

Figure 2.3: A staggered grid in one dimension of space.

the same precision on T at a given point using only one of these grids), it is also a common source of numerical difficulties, in particular if the two grids become weakly coupled (for example via an extra term in the equation).

The natural way to discretise (2.10) in the light of the above remarks is thus to use a staggered grid (i.e. only one of the above mentioned decoupled grids). This approach is illustrated on figure 2.3 in which $q = -\kappa \partial T/\partial x$ is defined on a dual grid, shifted half a grid point with respect to the original grid on which T is defined.

One can then easily introduce the following discretisation

$$q_{j+1/2} = -\kappa \left.\frac{\partial T}{\partial x}\right|_{j+1/2} \simeq -\kappa_{j+1/2}\,\frac{T_{j+1}-T_j}{\Delta x}\,, \tag{2.14}$$

$$\text{and} \qquad -\frac{\partial q}{\partial x} = \frac{\partial}{\partial x}\kappa(x)\frac{\partial T}{\partial x} \simeq \frac{\kappa_{j+1/2}\,(T_{j-1}-T_j)/\Delta x - \kappa_{j-1/2}\,(T_j-T_{j-1})/\Delta x}{\Delta x}\,. \tag{2.15}$$

In this expression κ may be given analytically and directly estimated on the dual grid, or it could be interpolated as $\kappa_{j+1/2} = (\kappa_j + \kappa_{j+1})/2$ if only known at the grid points. This would yield

$$\frac{\partial}{\partial x}\kappa(x)\frac{\partial}{\partial x} \simeq \frac{\kappa_{j+1/2}\,T_{j+1} - (\kappa_{j+1/2}+\kappa_{j-1/2})\,T_j + \kappa_{j-1/2}\,T_{j-1}}{\Delta x^2}\,. \tag{2.16}$$

This provides a natural way to discretise the equation and does not introduce any arbitrary choice between δ_0, δ_+, δ_-. It also prevents difficulties associated with grid decoupling.

It is interesting to note that the situation gets even worse in two dimensions of space as the heat flux becomes a vector

$$\boldsymbol{q} = -\kappa(\boldsymbol{x})\boldsymbol{\nabla} T\,, \qquad \frac{\partial T}{\partial t} = -\boldsymbol{\nabla}\cdot\boldsymbol{q}\,. \tag{2.17}$$

If we use the centered scheme but define all quantities at all points, we would then get four decoupled grids (see figure 2.4).

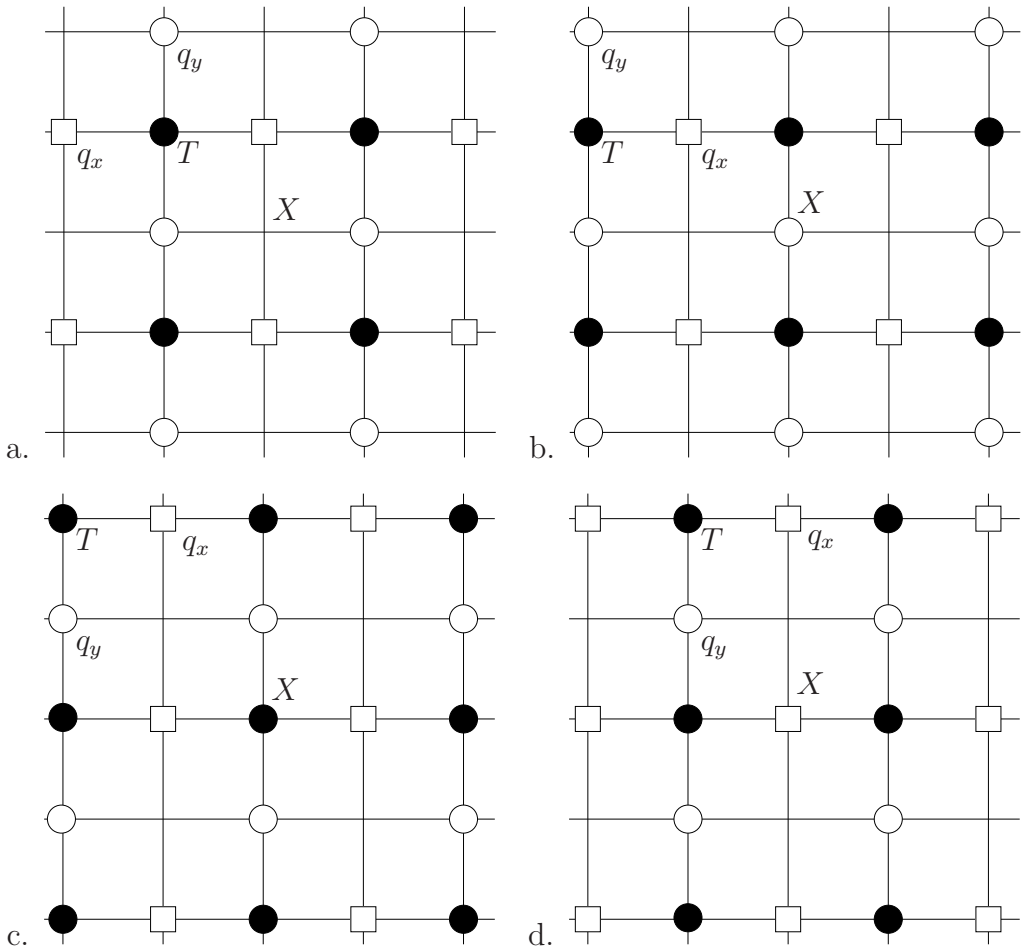

Figure 2.4: In two dimensions of space, four decoupled grid arise when using a centered scheme to discretise (2.17). The four grids are represented here using different symbols for T, $q_x = \boldsymbol{q} \cdot \boldsymbol{e}_x$ and $q_y = \boldsymbol{q} \cdot \boldsymbol{e}_y$. The symbol X is used to identify the same point on all four grids.

Similar grid decoupling occurs for other equations which can be reformulated in the form of a system of first order equations, for example the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}\,, \tag{2.18}$$

which can be rewritten in the form

$$\frac{\partial u}{\partial t} = -c\frac{\partial P}{\partial x}\,, \tag{2.19}$$

$$\frac{\partial P}{\partial t} = -c\frac{\partial u}{\partial x}\,, \tag{2.20}$$

a system which yields the same decoupling in space and also introduces a decoupling in time. The staggered grid formulation for this problem is

$$\frac{u^{n+1} - u^{n-1}}{2\Delta t} = -c\frac{P^n_{j+1} - P^n_{j-1}}{2\Delta x} \tag{2.21}$$

$$\frac{P^{n+1} - P^{n-1}}{2\Delta t} = -c\frac{u^n_{j+1} - u^n_{j-1}}{2\Delta x}\,. \tag{2.22}$$

One can eliminate P from the above expression

$$\frac{u^{n+1} - u^n}{\Delta t} = -c\frac{P^{n-1/2}_{j+1+2} - P^{n+1/2}_{j-1/2}}{\Delta x} \tag{2.23}$$

$$\frac{P^{n+1/2}_{j+1/2} - P^{n-1/2}_{j+1/2}}{\Delta t} = -c\frac{u^n_{j+1} - u^n_j}{\Delta x}\,, \tag{2.24}$$

and recover the classical scheme

$$\frac{u^{n+1}_j - 2u^n_j + u^{n-1}_j}{\Delta t^2} = c^2\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2}\,. \tag{2.25}$$

If a non staggered grid had been used, however, the same elimination would have provided the sparse stencil

$$\frac{u^{n+2}_j - 2u^n_j + u^{n-2}_j}{4\Delta t^2} = c^2\frac{u_{j+2} - 2u_j + u_{j-2}}{4\Delta x^4}\,. \tag{2.26}$$

2.4 Two dimensional problems

A straightforward way to handle two dimensional problems is to use one dimensional discretisation in two orthogonal directions, e.g.

$$\Delta u|_{i,j} = \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)\bigg|_{i,j} = \frac{u_{i-1,j} - 2\,u_{i,j} + u_{i+1,j}}{\Delta x^2} + \frac{u_{i,j-1} - 2\,u_{i,j} + u_{i,j+1}}{\Delta y^2} \tag{2.27}$$

This is the most natural way, but certainly not the only one.

An alternative scheme one could consider involves derivation accross the diagonals. This scheme is

$$\Delta u|_{i,j} = \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)\bigg|_{i,j} = \frac{(u_{i+1,j+1} - 2\,u_{i,j} + u_{i-1,j-1}) + (u_{i+1,j-1} - 2\,u_{i,j} + u_{i-1,j+1})}{\Delta x^2 + \Delta y^2} \quad (2.28)$$

It however introduces two decoupled grids in the form of a checker board (see `C2.1_2D_heati_vect.py`).

2.5 Boundary Conditions

Different types of boundary conditions can be considered.

2.5.1 The Dirichlet boundary contions

The value of u is then assumed to be fixed at the boundary (say $u = g$). This is straightforwardly implemented. For example for the heat equation we considered so far

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \kappa \frac{u_2^n - 2u_1^n + g}{\Delta x^2}\,. \quad (2.29)$$

2.5.2 The Neumann and Robin boundary conditions

An alternative form of boundary conditions often met is the *Neumann* boundary condition

$$\left.\frac{\partial u}{\partial x}\right|_0 = 0\,, \quad (2.30)$$

which is much less straightforward to implement. A reasonnable approach is to implement it by expressing this derivative using a finite difference scheme

$$\frac{u(1) - u(0)}{\Delta x} = 0\,. \quad (2.31)$$

However, if the boundary is located at 0 this will yield a first order aproximation, because a first order stencil was used at this stage (even if the rest of the domain is discretised using a second order formula).

One could alternatively consider

$$\frac{-3u_0 + 4u_1 - u_2}{2\Delta x} = 0 \quad \text{which yields} \quad u_0 = \frac{4}{3}u_1 - \frac{1}{3}u_2\,. \quad (2.32)$$

This does provide a second order expression, but at the cost of changing the width of the stencil at the boundary (the corresponding linear system would be more difficult to resolve).

An alternative approach is to use "ghost points", i.e. non physical computational points located on the other side of the boundary. Introducing a ghost point $u(-1)$ and using the second order formula

$$\left.\frac{\partial u}{\partial x}\right|_0 = \frac{u(1)-u(-1)}{2\Delta x} + \mathcal{O}(\Delta x^2) = 0\,, \tag{2.33}$$

and so for a homogeneous boundary condition, $u(-1) = u(1)$.

For the heat equation, this results in

$$\begin{aligned} u^{n+1}(0) &= u^n(0) + c\,[u^n(-1) - 2\,u^n(0) + u^n(1)] &(2.34)\\ &= u^n(0) + c\,[-2u^{\,n}(0) + 2\,u^n(1)]\,. &(2.35)\end{aligned}$$

Finally, one can use the dual grid introduced in the previous section and have the boundary on the dual grid. If the boundary is implemented at point $j = 1/2$, one gets

$$\left.\frac{\partial u}{\partial x}\right|_{1/2} = \frac{u(1)-u(0)}{\Delta x} + \mathcal{O}(\Delta x^2) = 0\,, \tag{2.36}$$

and so $u(0) = u(1)$.

This expression appears similar to (2.31), but should not be confused as Δx has been modified when shifting the boundary position. This restores second order accuracy.

In the case of *Robin* or *Fourier* boundary condition, of the form $\partial u/\partial x = \alpha u$, the procedure examplified above should be used, but rather in the form of (2.33) than (2.36), as both $\partial u/\partial x$ and u need to be known on the boundary.

2.6 Matrix formulation of finite differences

The scheme we considered for the heat equation

$$u_j^{n+1} - u_j^n = \frac{\kappa\Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n)\,, \tag{2.37}$$

can be written as

$$(u)^{n+1} = [A]\ (u)^n\ , \tag{2.38}$$

in which $[A]$ is a tri-diagonal matrix.

If we use an implicit scheme instead

$$u_j^{n+1} - u_j^n = \frac{\kappa\Delta t}{\Delta x^2}(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1})\,, \tag{2.39}$$

this matrix formulation will correspond to a linear system to be resolved.

The same sort of systems occur when solving for a potential problem

$$\Delta u = -f\,, \tag{2.40}$$

say for example the gravity potential $\Delta\phi = 4\pi G\mu$.

The matrix formulation is natural in one dimension of space. In two or three dimensions of space it becomes much more complicated. The construction of the resulting matrices is examplified in `C2.3_matrix.py` and uses the Kronecker product. It takes advantage of

$$\Delta_{2D} = Id \otimes \Delta_{1D} + \Delta_{1D} \otimes Id\,. \tag{2.41}$$

An essential point is that the matrices associated with finite differences stencils are always sparse (the matrix are composed essentially of zeros with a few non vanishing diagonals).

A direct consequence of this fact is that solving the linear system associated with the matrix will require $\mathcal{O}(N)$ operations in general (for a matrix of size $N \times N$). The inverse of the matrix however will be a full matrix of size $N \times N$. As a result the product of a vector with the inverse matrix would require $\mathcal{O}(N^2)$ operations. For this reason it is essential not to invert the matrix, but instead to solve for the linear system, even if this needs to be done at each timestep.

2.7 Compact Finite Differences

It is from formula in section 2.1 that in order to achieve higher order finite formula for a given derivative, one needs to involve more neighbouring points (i.e. a wider stencil). The extra degrees of freedom all to cancel more points in the Taylor expansion and thus to derive higher order expressions (this is clearly examplified by comparing expressions (2.7) and (2.8)). This also corresponds, in the formalism of section 2.2, to the use of higher degree polynomials.

While the formula are formaly of higher order, the approximation is not necessarily better. One is using a wider (non-local) stencil in order to get a finer estimate of a local quantity (the derivative at a point). If the function is not "smooth" at the scale of the grid this is not a sensible strategy.

An alternative strategy is to use a compact stencil ("compact" in the sens they only involve nearest neighbours) but introduce extra degrees of freedom by combining expressions of the unknown and not only the known quantities at these points. These expressions are also referred to as "Paddé" finite differences.

This is best illustrated on an example. Let us assume we want to derive an approximation of $\partial^2 u/\partial x^2 = -f$, using a stencil of the form

$$\alpha u_{j+1} + \beta u_j + \gamma u_{j-1} = a f_{j-1} + b f_j + c f_{j+1}\,. \tag{2.42}$$

Let us first consider the standard finite difference stencil (2.7)

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} - \frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4} + \mathcal{O}(\Delta x^4)\,. \tag{2.43}$$

$$\text{From} \quad \frac{\partial^2 u}{\partial x^2} = -f\,, \qquad \text{we get} \quad -\frac{\partial^4 u}{\partial x^4} = \frac{\partial^2 f}{\partial x^2} = \frac{f_{i+1} - 2f_i + f_{i-1}}{\Delta x^2} + \mathcal{O}(\Delta x^2)\,. \tag{2.44}$$

We can then propose the fourth order compact formula

$$\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2} = -f_j - \frac{f_{j+1}-2f_j+f_{j-1}}{12} + \mathcal{O}(\Delta x^4) \quad (2.45)$$

$$= -\frac{f_{i+1}+10f_i+f_{i-1}}{12} + \mathcal{O}(\Delta x^4)\,. \quad (2.46)$$

A similar approach can be used for the first order derivative $\partial u/\partial x = f$, and can for example provide

$$\frac{u_{j+1}-u_{j-1}}{2\Delta x} = \frac{1}{6}f_{j-1} + \frac{2}{3}f_j + \frac{1}{6}f_{j+1} + \mathcal{O}(\Delta x^4)\,. \quad (2.47)$$

2.8 Sources of physical errors

Let us now extend our heat equation to the case of a moving fluid

$$\frac{dT}{dt} = \frac{\partial T}{\partial t} + \boldsymbol{u}.\boldsymbol{\nabla}T = \kappa\Delta T\,, \quad (2.48)$$

or in one dimension of space

$$\frac{\partial T}{\partial t} + c\frac{\partial T}{\partial x} = \kappa\frac{\partial^2 T}{\partial x^2}\,. \quad (2.49)$$

If we discretise this equation using a second order scheme, we obtain

$$\frac{T^{n+1}-T^n}{\Delta t} = \kappa\frac{T^n_{j+1}-2T^n_j+T^n_{j-1}}{\Delta x^2} - c\left(\frac{T^n_{j+1}-T^n_{j-1}}{2\Delta x}\right)\,. \quad (2.50)$$

Let us introduce $\alpha = \kappa\Delta t/\Delta x^2$ and $\beta = c\Delta t/\Delta x$

$$T^{n+1}_j = T^n_j - \beta(T^n_{j+1}-T^n_{j-1}) + \alpha(T^n_{j-1}-2T^n_j+T^n_{j+1})\,. \quad (2.51)$$

We will now investigate the stability of this scheme in the purely advective case $\alpha = 0$,

$$\xi = 1 - \beta\left(\frac{e^{ik\Delta x}-e^{-ik\Delta x}}{2}\right) = 1 - \beta i\sin(k\Delta x)\,. \quad (2.52)$$

It follows that $|\xi| \geqslant 1$, and the scheme is thus unconditionnaly unstable. If α does not vanish, the diffusive term can stabilise the whole scheme.

The instability can be understood in physical terms by considering the modified equation

$$\frac{u^{n+1}-u^n}{\Delta t} - \frac{\Delta t}{2}\left(\frac{\partial^2 u}{\partial t^2}\right) = -c\left(\frac{u_{j+1}-u_{j-1}}{2\Delta x} - \frac{\Delta x^2}{6}\frac{\partial^3 u}{\partial x^3}\right)\,. \quad (2.53)$$

The truncation error is then

$$R_h = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + c\frac{\Delta x^2}{6}\frac{\partial^3 u}{\partial x^3} \quad \text{with} \quad \frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}\,. \quad (2.54)$$

Then at leading order, we have

$$\frac{\partial u}{\partial t} = c\frac{\partial u}{\partial x} - c^2\frac{\Delta t}{2}\frac{\partial^2 u}{\partial x^2}\,. \tag{2.55}$$

The truncation therefore introduces a reverse diffusion, which yields the instability (retrograd heat equation).

In order to achieve stability, a simple approach is to compensate explicitely for this term. This is the Lax-Wandroff scheme

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \frac{c^2\Delta t}{2}\,\frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\Delta x^2} = 0\,. \tag{2.56}$$

This scheme is stable if $c \leqslant \frac{\Delta x}{\Delta t}$.

Let us now consider the discretisation of the advective term using the δ_- stencil

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = -c\frac{u_j^n - u_{j-1}^n}{\Delta x}\,. \tag{2.57}$$

This scheme is first order in space and in time. We can investigate its stability

$$u_j^{n+1} = u_j^n - \beta(u_j^n - u_{j-1}^n)\,, \tag{2.58}$$

$$\xi = 1 - \beta(1 - e^{-ik\Delta x})\,, \tag{2.59}$$

and the scheme is thus stable if

$$0 < \beta < 1\,. \tag{2.60}$$

This condition requires $c > 0$ and $\Delta t < \Delta x/c$. The symmetric scheme δ_+, yields a similar condition with $c < 0$ and $\Delta t < \Delta x/|c|$.

A stable discretisation of the advective term therefore requires a discretisation involving the upwind (opposite to the direction of the velocity) differenciation. This is known as the upwind scheme.

2.9 Modified wave number

The concept of modified wave number is very useful in numerics as it allows to quantify the numerical distortion introduced on the physical problem.

This concept is directly related to the numerical diffusion encountered for he upwind scheme, it is the source of numerical diffusion, numerical dispersion and numerical anisotropy.

Let us consider a function of the form $f = e^{i(kx)}$, computing its discrete derivative via the centered scheme δ_0 yields

$$\frac{f_{j+1} - f_{j-1}}{2\Delta x} = i\,k'\,f_j\,, \tag{2.61}$$

where k' is a modified wave number.

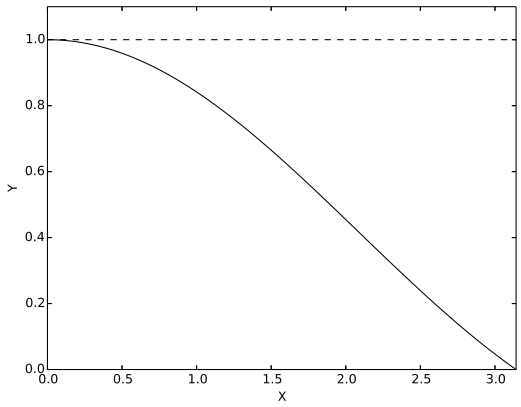

Figure 2.5: Modified wave number for the centered scheme.

One easily shows that

$$\frac{k'}{k} = \frac{\sin(k\Delta x)}{k\Delta x}\,. \tag{2.62}$$

Clearly $k'/k \to 1$ when $k\Delta x \to 0$. However, when the wavelength become comparable with the grid size, the effective wave number k' as seen by the stencil, significantly differs from the real wave number k (see figure 2.5).