# Homework LMECA2660: simulating convection

February 2024

We consider the 1-D convection equation with constant velocity $c$,

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = 0 \;, \tag{1}$$

in a periodic domain of period $L$. The initial condition is $u(x,0) = f(x)$. The solution at any time $t$ is then obtained (using the method of characteristics) as $u(x,t) = f(x - ct)$.

Using the Fourier transform, the initial condition can be represented as a sum of modes:

$$\begin{aligned} u(x,0) = f(x) \quad &= \quad \sum_{j=-\infty}^{\infty} \widehat{F}(k_j)\exp(\imath\, k_j\, x) \;, \\ \text{where} \quad \widehat{F}(k_j) \quad &= \quad \frac{1}{L}\int_{-L/2}^{L/2} f(x)\, \exp(-\imath k_j x)\, dx \end{aligned}$$

and with $k_j = \frac{2\pi}{L} j$ the so-called "wavenumber" of the mode. The exact solution at any time $t$ is thus obtained as

$$u(x,t) = \sum_{j=-\infty}^{\infty} \widehat{F}(k_j)\exp(\imath\, k_j\,(x - c\,t)) = \sum_{j=-\infty}^{\infty} \widehat{F}(k_j)\exp(\imath\,(k_j\, x - \omega_j\, t)) \;. \tag{2}$$

The propagation velocity is thus $c$ for all $k_j$. The corresponding "circular frequency" of the signal as perceived by an observer located a some fixed $x$ is $\omega_j = k_j\, c$ (and $\omega_j = 2\pi\, f_j$, with $f_j$ the frequency).

We first consider the case of a Gaussian initial condition in unbounded domain,

$$u(x,0) = U\, \exp\left(-\frac{x^2}{\sigma^2}\right) = \frac{Q}{\sqrt{\pi}\,\sigma}\, \exp\left(-\frac{x^2}{\sigma^2}\right) \tag{3}$$

with $Q = \int_{-\infty}^{\infty} u(x,0)\, dx = U\sqrt{\pi}\,\sigma$ the integral of the function. We also recall that the Fourier transform, $\widehat{u}(k,t)$, of the Gaussian is also a Gaussian:

$$\widehat{u}(k,0) = Q\, \exp\left(-\frac{k^2\,\sigma^2}{4}\right) \;. \tag{4}$$

The Gaussian function considered on a large periodic domain (i.e., with $L \gg \sigma$) is thus obtained as:

$$u(x,0) \simeq U\,\sqrt{\pi}\,\frac{\sigma}{L}\,\sum_{j=-\infty}^{\infty} \exp\left(-\frac{k_j^2\,\sigma^2}{4}\right)\,\exp(\imath\,k_j\,x)\ . \tag{5}$$

The larger the ratio $L/\sigma$, the better the approximation. The evolution in time is thus

$$u(x,t) \simeq U\,\sqrt{\pi}\,\frac{\sigma}{L}\,\sum_{j=-\infty}^{\infty} \exp\left(-\frac{k_j^2\,\sigma^2}{4}\right)\,\exp(\imath\,k_j\,(x - c\,t))\ . \tag{6}$$

- **Discretized periodic domain** :

  We consider fields that are discretized numerically using $N$ grid points, numbered from $i = 0$ to $i = N-1$, and with a uniform grid of size $h = \frac{L}{N}$; thus $x_i = -\frac{L}{2} + i\,h$. The discrete wavenumbers are thus limited to $0 \leq k_j\,h \leq \pi$.

  We hence also use the discrete Fourier series instead of the Fourier transform. For further details regarding the definitions and the links between those, see the "Reminder about Fourier transforms and Fourier series" at the end of this document.

  We first compare the spectral representation of the Gaussian function in an unbounded domain and the representation of the "periodized" Gaussian function obtained using the discrete Fourier transform. Using a FFT, obtain the coefficients of the discrete Fourier series for the Gaussian for the cases $\frac{L}{\sigma} = 4$ (small domain) and 16 (large domain), and for the finest resolution that will be used numerically later, $\frac{h}{\sigma} = \frac{1}{8}$. Compare them with the coefficients of the Fourier transform of the Gaussian function in unbounded domain by plotting the logarithm of their modulus as a function of $j$. For all the simulations to be done later, we will use the large domain with $\frac{L}{\sigma} = 16$.

  For the exact problem, the modes all move at the same velocity $c$. This is not the case when solving the problem numerically: the mode of wavenumber $k_j$ moves at the phase velocity $c_j^* = \frac{k_j^* h}{k_j h}\,c$ where $k_j^*$ is the "modified wavenumber" (see the lecture notes), and with $k_j^* h$ that depends on $k_j h$. The corresponding circular frequency is thus $\omega_j^* = k_j^*\,c$. For instance, using the second order centred scheme (E2) gives $k_j^* h = \sin(k_j h)$.

- **Partially decentered scheme** :

  1. Using Taylor series, obtain the partially decentered discretization for the convective term, $c\frac{\partial u}{\partial x}|_i$, using explicit finite differences involving $u_{i-2}$, $u_{i-1}$, $u_i$ and $u_{i+1}$. We will call it ED.

2. Determine the order and truncation error of scheme.
3. Using the modal analysis with $u_i(t) = \sum_j \widehat{U}_j(t) e^{ik_j x_i}$ and $\frac{d\widehat{U}_j}{dt} = \lambda_j \widehat{U}_j$, obtain the expression for $\lambda_j$ and for the modified dimensionless wavenumber $k^* h$ as a function of the dimensionless wavenumber $kh$.
4. Quantify the phase and/or amplitude error(s) by plotting the components of $\frac{k^* h}{\pi}$ as a function of $\frac{kh}{\pi}$.

- **Stability** :

  It is mandatory to also satisfy the stability constraints of the temporal integration scheme used; which will here be the classical Runge-Kutta 4 scheme (RK4C). It implies a limitation on the $CFL = \frac{c\Delta t}{h}$ number. Considering the eigenvalues $\lambda_j$ of the E2, E4, I4 and ED schemes for $0 \leq k_j\, h \leq \pi$, obtain the maximum allowable $CFL$ value.

  In order to make the comparison of the different spatial discretization schemes, all simulations will be performed using the same $CFL$ value (and properly rounded).

- **Numerical solution code** :

  Produce a C code:

  1. whose spatial discretization of the convective term is performed using either of the following schemes: E2, E4, I4 and ED.
  2. whose temporal integration is performed using RK4C.
  3. that outputs, at each time step, global diagnostics: $I_h^n = (h \sum_i u_i^n)/(\sigma\, U)$, $E_h^n = (h \sum_i (u_i^n)^2)/(\sigma\, U^2)$ and $R_h^n = (h \sum_i (u_i^n - u(x_i, t^n))^2)/(\sigma\, U^2)$ (measure of the global error).

  **Help:** To solve the periodic tridiagonal system required by the implicit I4 scheme, we use the *Thomas algorithm* that is described hereunder. The C code implementation of the algorithm is also provided.

- **Numerical simulation and analysis of the results** :

  Perform numerical simulations using each discretization scheme, and for three resolutions: $\frac{h}{\sigma} = \frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$ (thus corresponding to $N = \frac{L}{h} = 32$, 64 and 128).

  1. Plot the numerical solution $\frac{u_i^n}{U}$ and the analytical solution $\frac{u(x_i,t^n)}{U}$ from $-\frac{1}{2} \leq \frac{x}{L} \leq \frac{3}{2}$ (thus showing two periods) at the times $\frac{c\,t}{L} = \frac{1}{2}$ and 1. Comment.
  2. Plot the evolution of the global diagnostics, $I_h^n$, $E_h^n$ and $R_h^n$, as a function of $\frac{c\,t}{L}$. Comment.

3. Plot the global error achieved at $\frac{c\,t}{L} = \frac{1}{2}$ as a function of $\frac{h}{\sigma}$ in a log-log diagram. Determine the order of convergence of the different numerical methods used and comment.

We consider next the case where the grid spacing is not uniform. There is thus a mapping from the "numerical space" $\xi$ with $-\frac{L}{2} \leq \xi \leq \frac{L}{2}$ and with a uniform resolution $h$ (i.e., $\xi_i = -\frac{L}{2} + i\,h$) to the physical space $x$ with a non uniform resolution. This mapping is written as $x = g(\xi)$. The grid points in physical space are thus located at $x_i = g(\xi_i)$ with $-\frac{L}{2} \leq x_i \leq \frac{L}{2}$. The mapping used here is:

$$x = g(\xi) = \xi - a\,\frac{L}{2\pi}\,\sin\left(2\pi\frac{\xi}{L}\right) \ . \tag{7}$$

The derivative of the mapping is thus:

$$\frac{dx}{d\xi} = g'(\xi) = 1 - a\,\cos\left(2\pi\frac{\xi}{L}\right) \ . \tag{8}$$

The finest resolution in physical space is around $x = 0$ and is obtained as $\frac{dx}{d\xi} = 1 - a$. The coarsest resolution in physical space is around $x = \pm\frac{L}{2}$ and is obtained as $\frac{dx}{d\xi} = 1 + a$. The ratio between the finest and the coarsest resolutions is thus $\frac{1-a}{1+a}$.

It is easy to rewrite the transport equation in physical space with $c$ constant as a conservative equation for some function $v(\xi, t)$ in numerical space, and with some non uniform velocity $b(\xi)$:

$$\frac{\partial v}{\partial t} + \frac{\partial (b\,v)}{\partial \xi} = 0 \ . \tag{9}$$

Do it first! When done, perform numerical simulations for the case with $a = \frac{3}{5}$ (thus $\frac{1-a}{1+a} = \frac{1}{4}$ and using $N = 128$ grid points. For each scheme, plot the numerical solution $\frac{v_i^n}{U}$ as a function of $\frac{\xi_i}{L}$ at the times $\frac{c\,t}{L} = \frac{1}{2}$ and 1; and then also the corresponding $\frac{u_i^n}{U}$ as a function of $\frac{x_i}{L}$. Comment.

Finally, we consider the convection of a "wave packet": here, as simple model, a mode of specific wavenumber $k_p$ further modulated by a Gaussian:

$$u(x, 0) = U\,\cos(k_p\,x)\,\exp\left(-\frac{x^2}{\sigma^2}\right) \ . \tag{10}$$

We examine the case with $p = 12$. The wave packet can also be decomposed in Fourier modes: do it using the FFT and plot the spectrum (i.e., the modulus of the Fourier coefficients) as a function of $j$. Also compute the "group velocity" $c_g^*$ for each scheme and plot $c_g^*/c$ as a function of $j$.

We here investigate what happens to the wave packet when the resolution changes. We use again $a = \frac{3}{5}$ and $N = 128$ grid points. Plot, for each scheme, the numerical solution $\frac{u_i^n}{U}$ and the analytical solution $\frac{u(x_i, t^n)}{U}$ from $-\frac{1}{2} \leq \frac{x}{L} \leq \frac{3}{2}$ at the times $\frac{c\,t}{L} = \frac{1}{4}$ and

$\frac{1}{2}$. Comment.

**Reminder**: The evolution of a wave packet is governed by the so-called group velocity,

$$c_g^* = \frac{d\omega^*}{dk} = \frac{dk^*}{dk}\, c\,, \tag{11}$$

which is the velocity at which information can propagate (indeed, a pure mode $\cos(k_p\, x)$ cannot propagate any information). As such, it is of great importance in simulating physical problems. For instance, using the E2 scheme gives $c_g^* = \cos(k\, h)\, c$. As $\omega^*$ is then maximum at $k_m\, h = \frac{\pi}{2}$, the group velocity is zero there. For $k_m\, h < k\, h < \pi$, the group velocity is negative so that wave packets with wavenumbers mostly in this range will propagate upstream against the convection velocity; which is unphysical. Also note that, for any frequency $\omega^* < \omega^*(k_m) = \frac{c}{h}$, there are two wavenumbers that correspond to that frequency: one with positive group velocity and one with negative group velocity. The wavenumber with positive group velocity is a consistent approximation to a solution of the advection equation while the other one is spurious.

**Remarks and instructions**

1. This homework is a single person's job.
2. The program will be written in C.
3. The visualisations and FFTs can be performed using a software of your choice (e.g., Python, Matlab).
4. Provide clean and clear results. Provide nice and readable plots, with axes and caption (and legend when required).
5. A printed version (recto-verso) must be handed over.
6. A copy of your program and of your report must also be uploaded on **Moodle**. Instructions will follow.
7. The homework is due on Friday **March 15**.
8. **Note of warning**: A zero-tolerance policy will be applied regarding plagiarism. Systematic and automatic testing will be carried out (using the reports and the codes of the present and also the previous students). It is not forbidden to assist one another; but it must then be explicitly specified.

**Resolution of a tridiagonal and periodic system: extension of the Thomas algorithm**

We here present a summary of the Thomas algorithm. For further details, Google should be useful. The goal of the Thomas algorithm is to solve the system $Ax = q$ with:

$$A = \begin{bmatrix} a & c & & & & \\ b & a & c & & & \\ & b & a & c & & \\ & & \ddots & \ddots & \ddots & \\ & & & b & a & c \\ & & & & b & a \end{bmatrix} \tag{12}$$

This algorithm works in 2 steps: the *forward pass* whose aim is to simplify the equations, and the *backward pass* that effectively solves the system.

1. The forward pass simplifies the equations and sets them under the form

$$x_i + \tilde{a}_i x_{i+1} = \tilde{q}_i.$$

   The first equation becomes:

$$\tilde{a}_1 = \frac{c}{a} \qquad \tilde{q}_1 = \frac{q_1}{a} \tag{13}$$

   For the others, equation $(i)$ becomes $(i) - (i-1)b$:

$$\tilde{a}_i = \frac{c}{a - b\tilde{a}_{i-1}} \qquad \tilde{q}_i = \frac{q_i - b\tilde{q}_{i-1}}{a - b\tilde{a}_{i-1}} \tag{14}$$

   Indeed, by recurrence, one obtains

$$\left.\begin{array}{llllllll} (i-1) & x_{i-1} & + & \tilde{a}_{i-1}x_i & & & = & \tilde{q}_{i-1} \\ (i) & bx_{i-1} & + & ax_i & + & cx_{i+1} & = & q_i \end{array}\right\} \tag{15}$$

$$(i) - (i-1)\,b \quad \Rightarrow \quad x_i + \frac{c}{a - b\tilde{a}_{i-1}} x_{i+1} = \frac{q_i - b\tilde{q}_{i-1}}{a - b\tilde{a}_{i-1}} \tag{16}$$

2. Backward pass. Now that the system is upper triangular, we are able to solve it very efficiently. For the last equation:

$$x_n = \tilde{q}_n \tag{17}$$

   For the other equations, we proceed from bottom to top:

$$x_i = \tilde{q}_i - \tilde{a}_i x_{i+1} \tag{18}$$

However, the system we here wish to solve is a periodic problem. The matrix is then:

$$A = \begin{bmatrix} a & c & & & & b \\ b & a & c & & & \\ & b & a & c & & \\ & & \ddots & \ddots & \ddots & \\ & & & b & a & c \\ c & & & & b & a \end{bmatrix}, \tag{19}$$

Therefore, we are not able to directly use *Thomas Algorithm* as detailed above. However, let's consider the matrix $A_c = A\,(1 : n-1, 1 : n-1)$, i.e. the $A$ matrix without the last row and the last column. Such a matrix satisfies the structure required by the algorithm. The problem then becomes:

$$A_c x_c = q_c - \begin{bmatrix} b \\ \vdots \\ c \end{bmatrix} x_n \tag{20}$$

$$b x_{n-1} + a x_n + c x_1 = q_n \tag{21}$$

To solve the first equation, let's assume that $x_c$ has the form:

$$x_c = x^{(1)} + x^{(2)} x_n \tag{22}$$

In that case, $x_1$, $x_2$ are given by (linearity of the system):

$$A_c x^{(1)} = q_c \qquad A_c x^{(2)} = - \begin{bmatrix} b \\ \vdots \\ c \end{bmatrix}. \tag{23}$$

This allows to solve the last equation:

$$x_n = \frac{q_n - c x_1^{(1)} - b x_{n-1}^{(1)}}{a + c x_1^{(2)} + b x_{n-1}^{(2)}}. \tag{24}$$

**Reminder about Fourier transforms and Fourier series**

The Fourier transform $\widehat{f}(k)$ of a function $f(x)$ is defined as:

$$\widehat{f}(k) = \mathcal{F}\left(f(x)\right) \;=\; \int_{-\infty}^{\infty} f(x)\, \exp\left(-\imath k x\right) dx \;,$$

while the inverse transform is defined as:

$$f(x) = \mathcal{F}^{-1}\left(\widehat{f}(k)\right) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(k)\, \exp\left(\imath k x\right) dk \;.$$

We easily verify that:

$$\mathcal{F}\left(f(x-x_0)\right) \;=\; \widehat{f}(k)\;\exp\left(-\imath k x_0\right) \;.$$

Onde can also demonstrate that:

$$\mathcal{F}\left(\exp\left(-\frac{x^2}{\sigma^2}\right)\right) \;=\; \sqrt{\pi}\,\sigma\;\exp\left(-\frac{k^2\,\sigma^2}{4}\right) \;.$$

A periodic function $f(x)$, of period $L$, may be represented by a Fourier series:

$$f(x) \;=\; \sum_{j=-\infty}^{\infty} \widehat{F}(k_j) \exp(\imath k_j x) \;,$$

$$\text{where} \quad \widehat{F}(k_j) \;=\; \frac{1}{L}\int_{-L/2}^{L/2} f(x)\, \exp(-\imath k_j x)\, dx$$

and with $k_j = \frac{2\pi}{L} j$. We easily verify that the Fourier series of the function shifted in space by $x_0$, $f(x-x_0)$, is simply $\widehat{F}(k_j)\;\exp\left(-\imath k_j x_0\right)$.

Finally, a discrete and periodic function of period $L$, with $x_i = -\frac{L}{2} + i\,h$ ($i = 0, 1, 2, \ldots, N-1$, where $N$ is even, and $h = \frac{L}{N}$) and $f_i = f(x_i)$, can be represented by a discrete Fourier series:

$$f_i = f(x_i) \;=\; \sum_{j=-N/2}^{N/2} \widehat{F}(k_j) \exp(\imath k_j x_i) \;,$$

$$\text{where} \quad \widehat{F}_j = \widehat{F}(k_j) \;=\; \frac{1}{N}\sum_{i=0}^{N-1} f(x_i)\, \exp(-\imath k_j x_i) \;.$$

Here again, the discrete Fourier series of the function shifted in space by $x_0$, $f(x_i - x_0)$, is simply: $\widehat{F}(k_j)\;\exp\left(-\imath k_j x_0\right)$.

The coefficients $\widehat{F}(k_j)$ are typically obtained using the Fast Fourier Transform (FFT). If the function $f(x_i)$ is real-valued, the coefficients are complex conjugates: $\widehat{F}(-k_j) = \widehat{F}_r(-k_j) + \imath\, \widehat{F}_\imath(-k_j) = \left(\widehat{F}(k_j)\right)^* = \widehat{F}_r(k_j) - \imath\, \widehat{F}_\imath(k_j)$. We then also have:

$$f(x_i) \;\; = \;\; \widehat{F}_r(k_0) + \sum_{j=1}^{N/2-1} 2\left[\widehat{F}_r(k_j)\cos(k_j x_i) - \widehat{F}_\imath(k_j)\sin(k_j x_i)\right] + 2\widehat{F}_r(k_{N/2})\cos(k_{N/2} x_i) \; .$$

The first term (i.e., the "zero mode") corresponds to the mean value of the function. The last term (i.e. the "flip-flop mode") corresponds to the highest wavenumber mode (i.e., $k_{N/2}h = \pi$) and is taken as purely real (its imaginary part is set to zero). We thus have $N$ discrete and real values of the function $(f_0, f_1, \ldots, f_{N-1})$, and $N$ discrete and real values for the reconstruction of the Fourier coefficients: $1 + 2 * (N/2 - 1) + 1 = N$

By extension, a non periodic function, but evaluated on an interval $L$ which is large enough for the function to be considered periodic, may be represented approximately on this interval $L$ by using the Fourier transform $\widehat{f}(k)$ obtained in an unbounded domain:

$$f(x) \;\; \simeq \;\; \frac{1}{L}\sum_{j=-\infty}^{\infty} \widehat{f}(k_j)\exp(\imath k_j x) \; .$$

For example, for the Gaussian function considered on a periodic domain with $L \gg \sigma$, we obtain:

$$f(x) \;\; \simeq \;\; U\sqrt{\pi}\,\frac{\sigma}{L}\sum_{j=-\infty}^{\infty} \exp\left(-\frac{k_j^2\,\sigma^2}{4}\right)\exp(\imath\, k_j\, x) \; .$$

The larger the ratio $L/\sigma$, the better the approximation.

For the discrete version, we then obtain:

$$f_i = f(x_i) \;\; \simeq \;\; \frac{1}{L}\sum_{j=-N/2}^{N/2} \widehat{f}(k_j)\exp(\imath k_j x_i) \; ,$$

and thus, for the Gaussian function considered on a periodic domain,

$$f_i = f(x_i) \;\; \simeq \;\; U\sqrt{\pi}\,\frac{\sigma}{L}\sum_{j=-N/2}^{N/2} \exp\left(-\frac{k_j^2\,\sigma^2}{4}\right)\exp(\imath\, k_j\, x_i) \; .$$