Finite Element methods for hyperbolic systems

Eric Sonnendrücker
Max-Planck-Institut für Plasmaphysik
und
Zentrum Mathematik, TU München

Lecture notes
Wintersemester 2014-2015

January 19, 2015

Contents

Chapter 1

Introduction

Hyperbolic systems arise naturally from the conservation laws of physics. Writing down the conservation of mass, momentum and energy yields a system of equations that needs to be solved in order to describe the evolution of the system. In this lecture we will introduce the classical methods for numerically solving such systems. Up to a few years ago these were essentially finite difference methods and finite volume methods. But in the last decades a new class of very efficient and flexible method has emerged, the Discontinuous Galerkin method, which shares some features both with Finite Volumes and Finite Elements.

In 1D the systems of conservation laws we consider have the form

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{u})}{\partial x} = 0,$$

where $\mathbf{u}$ is a vector of unknown values. This can be written also

$$\frac{\partial \mathbf{u}}{\partial t} + A(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0,$$

where $A(\mathbf{u})$ is the Jacobian matrix with components $((\frac{\partial F_i}{\partial u_j}))_{i,j}$. The system is called *hyperbolic* if for all $\mathbf{u}$ the matrix A has only real eigenvalues and is diagonalisable. It is called *strictly hyperbolic* if all eigenvalues are distinct.

Examples:

- 1D Maxwell's equation

$$\begin{aligned} \frac{\partial E}{\partial t} + \frac{\partial B}{\partial x} &= J \\ \frac{\partial B}{\partial t} + \frac{\partial E}{\partial x} &= 0 \end{aligned}$$

- 1D Euler equations

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho u \\ \rho u^2 + p \\ Eu + pu \end{pmatrix} = 0,$$

  where ρ, u and E are the density, velocity and energy density of the gas and p is the pressure which is a known function of ρ.

The idea behind all numerical methods for hyperbolic systems is to use the fact that the system is locally diagonalisable, with real eigenvalues, and thus can be reduced to a set of scalar equations. For this reason, before going to systems it will be useful to first understand the scalar case and then see how it can be extended to systems by local diagonalization.

The first part of the lecture will be devoted to the linear case, starting with the scalar case which boils down to linear advection for which the core methods will be first introduced. This is fairly straightforward.

Many additional problems arise in the nonlinear case. Indeed in this case even starting from a smooth initial condition discontinuities can appear in the solution. In this case the concept of weak solutions need to be introduced and there can be several solutions only one of which is physical. We thus need a criterion to find the physical solution and numerical schemes that capture the physical solution. The notion of conservativity plays an essential role there. These will be addressed in the second part.

Chapter 2

1D linear advection

2.1 Finite Difference schemes for the advection equation

We consider first the linear 1D advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \qquad \text{pour } x \in [0, L],\ t \geq 0. \tag{2.1}$$

Let us assume for simplicity that the boundary conditions are periodic. This means that u and all its derivatives are periodic of period L. We have in particular $u(0) = u(L)$. The constant a is given. As the problem is time dependent we also need an initial condition $u(x, 0) = u_0(x)$.

2.1.1 Obtaining a Finite Difference scheme

We first consider a uniform mesh of the 1D computational domain, i.e. of the interval $[a, b]$ where we want to compute the solution, see Figure 2.1. The cell size or space step

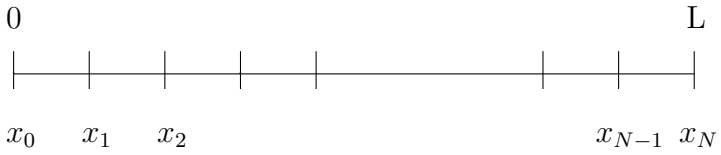

Figure 2.1: Uniform mesh of $[a, b]$

is defined by $\Delta x = \frac{L}{N}$ where N is the number of cells in the mesh. The coordinates of the grid points are then defined by $x_i = x_0 + i\Delta x$. We then need a time step Δt and we will compute approximations of the solution at discrete times $t_n = n\Delta t$, $n \in \mathbb{N}$. As we assume the solution to be periodic of period L it will be defined by its values at x_i for $0 \leq i \leq N-1$ and we shall have $u(x_N, t_n) = u(x_0, t_n)$.

We shall denote by $u_j^n = u(x_j, t_n)$.

2.1.2 The first order explicit upwind scheme

A Finite Difference scheme is classically obtained by approximating the derivatives appearing in the partial differential equation by a Taylor expansion up to some given order which will give the order of the scheme. As we know only the values of the unknown

function at the grid points, we use Taylor expansion at different grid points and linearly combine them so as to eliminate all derivatives up to the needed order.

The same can be done for the time discretisation. For an approximation of order 1 in space and time, we can simply write

$$\frac{\partial u}{\partial t}(x_j, t_n) = \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t} + O(\Delta t), \tag{2.2}$$

$$\frac{\partial u}{\partial x}(x_j, t_n) = \frac{u(x_j, t_n) - u(x_{j-1}, t_n)}{\Delta x} + O(\Delta x). \tag{2.3}$$

Denoting by u_j^n, the approximation of the solution at point x_j and time t_n and using the above formulas for the approximation of the partial derivatives we get the following approximation (2.1) at point x_j and time t_n:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0. \tag{2.4}$$

We thus obtain the following explicit formula which enables to compute u_j^{n+1} in function of the values of u at time t_n and points x_{j-1}, x_j and x_{j-1} :

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{\Delta x}(u_j^n - u_{j-1}^n). \tag{2.5}$$

Denote by U^n the vector of $\mathbb{R}^N$ whose components are $u_0^n, \ldots, u_{N-1}^n$ and

$$A = \begin{pmatrix} (1 - \frac{a\Delta t}{\Delta x}) & 0 & & \frac{a\Delta t}{\Delta x} \\ \frac{a\Delta t}{\Delta x} & \ddots & \ddots & \\ & \ddots & \ddots & 0 \\ 0 & & \frac{a\Delta t}{\Delta x} & (1 - \frac{a\Delta t}{\Delta x}) \end{pmatrix}.$$

The terms at the end of the first line comes from the periodic boundary conditions. We use that $u_{-1}^n = u_{N-1}^n$ and $u_N^n = u_0^n$. Except on the two diagonals all the terms vanish

In this case the scheme (2.5) can be written in matrix form

$$U^{n+1} = AU^n.$$

2.1.3 The first order upwind implicit scheme

When using an uncentered difference scheme in the other direction for the time derivative, we get

$$\frac{\partial u}{\partial t}(x_j, t_n) = \frac{u(x_j, t_n) - u(x_j, t_{n-1})}{\Delta t} + O(\Delta t), \tag{2.6}$$

We use the same finite difference approximation for the space derivative. We then get the following formula

$$u_j^n + a\frac{\Delta t}{\Delta x}(u_j^n - u_{j-1}^n) = u_j^{n-1}. \tag{2.7}$$

In this case the u_j^n are defined implicitly from the u_j^{n-1} as solutions of a linear system. This is why this scheme is called implicit.

Denote by B the matrix of the linear system:

$$B = \begin{pmatrix} (1+\frac{a\Delta t}{\Delta x}) & 0 & & -\frac{a\Delta t}{\Delta x} \\ -\frac{a\Delta t}{\Delta x} & \ddots & \ddots & \\ & \ddots & \ddots & 0 \\ 0 & & -\frac{a\Delta t}{\Delta x} & (1+\frac{a\Delta t}{\Delta x}) \end{pmatrix}.$$

The term at the end of the first line comes from the periodic boundary conditions. We use that $u^n_{-1} = u^n_{N-1}$ and $u^n_N = u^n_0$. The terms not on the two diagonals vanish.

Going now from time step n to $n+1$ the implicit scheme in matrix form becomes

$$BU^{n+1} = U^n.$$

2.1.4 The method of lines

As we saw, the time discretisation can be performed by finite differences as for the space discretisation. However it is generally more convenient to separate the space and time discretisation for a better understanding. The method of lines consists in applying only a discretisation scheme in space first (this can be Finite Differences or any other scheme). Then one obtains a system of ordinary differential equations of the form

$$\frac{dU}{dt} = \mathcal{A}U,$$

where $U(t)$ is the vector whose components are $u_i(t)$ the unknown values at the grid point at any time t. Then one can use any Ordinary Differential Equation (ODE) solver for the time discretisation. For example using an explicit Euler method with the upwind method in space yields the previous explicit upwind scheme and when we use an implicit Euler method we get the implicit upwind scheme.

2.1.5 Convergence of finite difference schemes

So as to include explicit and implicit schemes, we consider a linear scheme in the following generic matrix form

$$LU^{n+1} = MU^n, \tag{2.8}$$

where the matrices L and M, are normalised such that the coefficient of u^n_j, is 1.

Let us denote by

$$V(t_n) = \begin{pmatrix} u(x_0, t_n) \\ \vdots \\ u(x_{N-1}, t_n), \end{pmatrix}$$

where u is the exact solution of the Partial Differential Equation (PDE) that is approximated by the scheme (2.8).

Definition 1 *The scheme (2.8) is called consistent of order (q, p) if*

$$\|LV(t_{n+1}) - MV(t_n)\| = \Delta t\, O(\Delta t^q + \Delta x^p).$$

Proposition 1 *If the exact solution is of class C^3, the schemes (2.5) and (2.7) are consistent of order 1 in t and in x.*

Proof. Let's start with (2.5). Using the Taylor formulas (2.2) and (2.3), we obtain for the i^{th} line of $V(t_{n+1}) - AV(t_n)$,

$$\begin{aligned}(V(t_{n+1}) - AV(t_n))_i &= u(x_i, t_{n+1}) - u(x_i, t_n) + \frac{a\Delta t}{\Delta x}(u(x_i, t_n) - u(x_{i-1}, t_n)) \\ &= \Delta t \left(\frac{\partial u}{\partial t}(x_i, t_n) + O(\Delta t) + a\frac{\partial u}{\partial x}(x_i, t_n) + O(\Delta x) \right).\end{aligned}$$

The result follows as u is a solution of our equation. We thus get the (1,1) consistency of this scheme.

Then for (2.7), we use (2.6) and (2.3). The i^{th} line of $BV(t_n) - V(t_{n-1})$,

$$\begin{aligned}(BV(t_n) - V(t_{n-1}))_i &= u(x_i, t_n) - u(x_i, t_{n-1}) + \frac{a\Delta t}{\Delta x}(u(x_i, t_n) - u(x_{i-1}, t_n)) \\ &= \Delta t \left(\frac{\partial u}{\partial t}(x_i, t_n) + O(\Delta t) + a\frac{\partial u}{\partial x}(x_i, t_n) + O(\Delta x) \right).\end{aligned}$$

The result follows as u is a solution of the equation. We thus get the (1,1) consistency of this scheme. ■

Definition 2 *The scheme (2.8) is called stable for some given norm $\|.\|$ if there exist constants K and τ independent of Δt such that*

$$\|U^n\| \leq K\|U^0\| \quad \forall \Delta t \text{ such that } 0 < \Delta t < \tau.$$

Proposition 2 *The scheme (2.5) is stable for the L^∞ norm provided*

$$\frac{a\Delta t}{\Delta x} \leq 1.$$

This condition is called the **Courant-Friedrichs-Lewy or CFL condition**.

Proof. Consider the scheme (2.5). Grouping the terms in u_{j-1}^n, u_j^n and u_{j+1}^n, this scheme (2.5) becomes

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

As, a, Δt and Δx are positive, if $\frac{a\Delta t}{\Delta x} \leq 1$ the two factors of u_{j-1}^n, u_j^n are positive and equal to their absolute value. Hence

$$|u_j^{n+1}| \leq \frac{a\Delta t}{\Delta x}|u_{j-1}^n| + (1 - \frac{a\Delta t}{\Delta x})|u_j^n| \leq \big(\frac{a\Delta t}{\Delta x} + (1 - \frac{a\Delta t}{\Delta x})\big) \max_j |u_j^n|,$$

and so

$$\max_j |u_j^{n+1}| \leq \max_j |u_j^n|,$$

from which it follows that $\max_j |u_j^n| \leq \max_j |u_j^0|$ for all n, which yields the L^∞ stability. ■

von Neumann stability analysis. Due to the fact that the discrete Fourier transform conserves the L^2 norm because of the discrete Plancherel inequality and that it diagonalises the Finite Difference operators (provided the original PDE has constant coefficients), it is particularly well adapted for studying the L^2 stability. The von Neumann analysis consists in applying the discrete Fourier transform to the discretised equation. To make this more precise we first recall the definition and main properties of the discrete Fourier transform.

Let P be the symmetric matrix formed with the powers of the n^{th} roots of unity the coefficients of which are given by $P_{jk} = \frac{1}{\sqrt{n}} e^{\frac{2i\pi jk}{n}}$. Denoting by $\omega_n = e^{\frac{2i\pi}{n}}$, we have $P_{jk} = \frac{1}{\sqrt{n}} \omega_n^{jk}$.

Notice that the columns of P, denoted by P_i, $0 \leq i \leq n-1$ are the vectors X_i normalised so that $P_i^* P_j = \delta_{i,j}$. On the other hand the vector X_k corresponds to a discretisation of the function $x \mapsto e^{-2i\pi kx}$ at the grid points $x_j = j/n$ of the interval $[0, 1]$. So the expression of a periodic function in the base of the vectors X_k is thus naturally associated to the Fourier series of a periodic function.

Definition 3 *Discrete Fourier Transform.*

- *The* ***dicrete Fourier transform*** *of a vector $x \in \mathbb{C}^n$ is the vector $y = P^* x$.*
- *The* ***inverse discrete Fourier transform*** *of a vector $y \in \mathbb{C}^n$ is the vector $x = P^{*-1} x = Px$.*

Lemma 1 *The matrix P is unitary and symmetric,* i.e. *$P^{-1} = P^* = \bar{P}$.*

Proof. We clearly have $P^T = P$, so $P^* = \bar{P}$. There remains to prove that $P\bar{P} = I$. But we have

$$(P\bar{P})_{jk} = \frac{1}{n} \sum_{l=0}^{n-1} \omega^{jl} \omega^{-lk} = \frac{1}{n} \sum_{l=0}^{n-1} e^{\frac{2i\pi}{n} l(j-k)} = \frac{1}{n} \frac{1 - e^{\frac{2i\pi}{n} n(j-k)}}{1 - e^{\frac{2i\pi}{n}(j-k)}},$$

and so $(P\bar{P})_{jk} = 0$ if $j \neq k$ and $(P\bar{P})_{jk} = 1$ if $j = k$. ∎

Corollary 1 *Let $F, G \in \mathbb{C}^n$ and denote by $\hat{F} = P^* F$ and $\hat{G} = P^* G$, their discrete Fourier transforms. Then we have*

- *the discrete Parseval identity:*

$$(F, G) = F^T \bar{G} = \hat{F}^T \bar{\hat{G}} = (\hat{F}, \hat{G}), \tag{2.9}$$

- *The discrete Plancherel identity:*

$$\|F\| = \|\hat{F}\|, \tag{2.10}$$

where $(.,.)$ and $\|.\|$ denote the usual euclidian dot product and norm in $\mathbb{C}^n$.

Proof. The dot product in $\mathbb{C}^n$ of $F = (f_1, \ldots, g_n)^T$ and $G = (g_1, \ldots, g_n)^T$ is defined by

$$(F, G) = \sum_{i=1}^{N} f_i \bar{g}_i = F^T \bar{G}.$$

Then using the definition of the inverse discrete Fourier transform, we have $F = P\hat{F}$, $G = P\hat{G}$, we get

$$F^T\bar{G} = (P\hat{F})^T\overline{P\hat{G}} = \hat{F}^T P^T \bar{P}\bar{\hat{G}} = \hat{F}^T\bar{\hat{G}},$$

as $P^T = P$ and $\bar{P} = P^{-1}$. The Plancherel identity follows from the Parseval identity by taking $G = F$. ■

Remark 1 *The discrete Fourier transform is defined as a matrix-vector multiplication. Its computation hence requires a priori n^2 multiplications and additions. But because of the specific structure of the matrix there exists a very fast algorithm, called Fast Fourier Transform (FFT) for performing it in $O(n\log_2 n)$ operations. This makes it particularly interesting for many applications, and many fast PDE solvers make use of it.*

Let us now consider the generic matrix form of the Finite Difference scheme introduced above:

$$LU^{n+1} = MU^n.$$

Note that on a uniform grid if the PDE coefficients do not explicitly depend on x the scheme is identical at all the grid points. This implies that L and M have the same coefficients on any diagonal including the periodicity. Such matrices, which are of the form

$$C = \begin{pmatrix} c_0 & c_1 & c_2 & \dots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & & c_{n-3} \\ \vdots & & & \ddots & \vdots \\ c_1 & c_2 & c_3 & \dots & c_0 \end{pmatrix}$$

with $c_0, c_1, \dots, c_{n-1} \in \mathbb{R}$ are called *circulant*.

Proposition 3 *The eigenvalues of the circulant matrix C are given by*

$$\lambda_k = \sum_{j=0}^{n-1} c_j \omega^{jk}, \tag{2.11}$$

where $\omega = e^{2i\pi/n}$.

Proof. Let J be the circulant matrix obtained from C by taking $c_1 = 1$ and $c_j = 0$ for $j \neq 1$. We notice that C can be written as a polynomial in J

$$C = \sum_{j=0}^{n-1} c_j J^j.$$

As $J^n = I$, the eigenvalues of J are the n-th roots of unity that are given by $\omega^k = e^{2ik\pi/n}$. Looking for X_k such that $JX_k = \omega^k X_k$ we find that an eigenvector associated to the eigenvalue λ_k is

$$X_k = \begin{pmatrix} 1 \\ \omega^k \\ \omega^{2k} \\ \vdots \\ \omega^{(n-1)k} \end{pmatrix}.$$

We then have that

$$CX_k = \sum_{j=0}^{n-1} c_j J^j X_k = \sum_{j=0}^{n-1} c_j \omega^{jk} X_k,$$

and so the eigenvalues of C associated to the eigenvectors X_k are

$$\lambda_k = \sum_{j=0}^{n-1} c_j \omega^{jk}.$$

■

Proposition 4 *Any circulant matrix C can be written in the form $C = P\Lambda P^*$ where P is the matrix of the discrete Fourier transform and Λ is the diagonal matrix of the eigenvalues of C. In particular all circulant matrices have the same eigenvectors (which are the columns of P), and any matrix of the form $P\Lambda P^*$ is circulant.*

Corollary 2 *We have the following properties:*

- *The product of two circulant matrix is circulant matrix.*
- *A circulant matrix whose eigenvalues are all non vanishing is invertible and its inverse is circulant.*

Proof. The key point is that all circulant matrices can be diagonalized in the same basis of eigenvectors. If C_1 and C_2 are two circulant matrices, we have $C_1 = P\Lambda_1 P^*$ and $C_2 = P\Lambda_2 P^*$ so $C_1 C_2 = P\Lambda_1\Lambda_2 P^*$.

If all eigenvalues of $C = P\Lambda P^*$ are non vanishing, Λ^{-1} is well defined and

$$P\Lambda P^* P\Lambda^{-1} P^* = I.$$

So the inverse of C is the circulant matrix $P\Lambda^{-1}P^*$. ■

Now applying the discrete Fourier transform to our generic scheme yields:

$$P^*LU^{n+1} = P^*MU^n \quad \Leftrightarrow \quad P^*LPP^*U^{n+1} = P^*MPP^*U^n$$

which is equivalent to

$$\Lambda_L \hat{U}^{n+1} = \Lambda_M \hat{U}^n,$$

where Λ_L and Λ_M are the diagonal matrices containing the eigenvalues of the circulant matrices M and L which are given explicitly from the matrix coefficients. It follows because $\|\hat{U}\| = \|U\|$ for any vector U that the scheme is L^2 stable if and only if $\max_i \frac{|\lambda_{M,i}|}{|\lambda_{L,i}|} \leq 1$, where $\lambda_{M,i}$ and $\lambda_{L,i}$ are the eigenvalues of M and L.

Let us now apply this technique to the scheme (2.7):

Proposition 5 *The scheme (2.7) is stable for the L^2 norm for all strictly positive values of Δx and Δt.*

Proof. Let us denote by $\alpha = a\frac{\Delta t}{\Delta x}$. We notice that matrix B is circulant with $c_0 = 1+\alpha$, $c_{n-1} = -\alpha$, the other c_i being 0.

The eigenvalues of B thus are $\lambda_k = c_0 + c_{n-1}\omega^{-2ik\pi/n}$. Which implies that

$$\Re\lambda_k = 1 + \alpha(1 - \cos\frac{2k\pi}{n}) \geq 1,$$

as $\alpha \geq 0$. It follows that all eigenvalues of B have a modulus larger or equal to 1, which implies that B is invertible Moreover the eigenvalues of its inverse all have modulus less or equal to 1, which implies the L^2 stability of the scheme. ■

Proposition 6 *The explicit centred scheme second order in space and first order in time:*

$$u_j^{n+1} = u_j^n - \frac{a\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n).$$

is unstable in L^2.

Proof. Let us denote by $\alpha = a\frac{\Delta t}{\Delta x}$. The first order in time centred scheme becomes in matrix form $U^{n+1} = AU^n$ where A is the circulant matrix with three non vanishing diagonals corresponding to $c_0 = 1$, $c_1 = -c_{n-1} = \frac{\alpha}{2}e^{\frac{2i\pi k}{n}}$. Hence its eigenvalues are $\lambda_k = 1 - \frac{\alpha}{2}(e^{\frac{2i\pi k}{n}} - e^{-\frac{2i\pi k}{n}}) = 1 - i\alpha\sin\frac{2k\pi}{n}$ so that $|\lambda_k| > 1$ for all k such that $\sin\frac{2k\pi}{n} \neq 0$ if $\alpha \neq 0$. Hence the scheme is unstable. ■

Theorem 1 (Lax) *The scheme (2.8) is convergent if it is stable and consistent.*

Proof. Let $V(t_n)$ be the vector whose components are the exact solution at the grid points at time t_n. The, as the scheme is consistent, we have

$$LV(t_{n+1}) = MV(t_n) + \Delta tO(\Delta t^q + \Delta x^p).$$

Note $E^n = U^n - V(t_n)$ the vector containing the errors at each point at time t_n, then as $LU^{n+1} = MU^n$, we have $LE^{n+1} = ME^n + \Delta tO(\Delta t^q + \Delta x^p)$. Hence

$$\begin{aligned} E^{n+1} &= L^{-1}ME^n + \Delta tK_1(\Delta t^q + \Delta x^p), \\ &= L^{-1}M(E^{n-1} + \Delta tK_1(\Delta t^q + \Delta x^p)) + \Delta tK_1(\Delta t^q + \Delta x^p), \\ &= (L^{-1}M)^{n+1}E^0 + (1 + \cdots + (L^{-1}M)^n)\Delta tK_1(\Delta t^q + \Delta x^p). \end{aligned}$$

Hence

$$\|E^{n+1}\| \leq \|(L^{-1}M)^{n+1}E^0\| + \|(1 + \cdots + (L^{-1}M)^n)\Delta tK_1(\Delta t^q + \Delta x^p)\|. \tag{2.12}$$

The stability implies that for any initial condition U^0, as $U^n = (L^{-1}M)^nU^0$, we have

$$\|(L^{-1}M)^nU^0\| \leq K\|U^0\|,$$

which means that $\|(L^{-1}M)^n\| \leq K$ for all n. Plugging this into (2.12), we obtain:

$$\|E^{n+1}\| \leq K\|E^0\| + nKK_1\Delta t(\Delta t^q + \Delta x^p).$$

Then as on the one hand E^0 taking as an initial in the scheme $U^0 = V(0)$, and on the other hand $n\Delta t \leq T$ the maximal considered time, we have

$$\|E^{n+1}\| \leq KK_1T(\Delta t^q + \Delta x^p),$$

whence convergence. ■

Corollary 3 *If the exact solution is of class C^3, the schemes (2.5) and (2.7) converge.*

Proof. This follows immediately from the Lax theorem by applying propositions 1 and 2. ■

2.1.6 High-order time schemes

When working with linear homogeneous equations with no source term, the simplest way to derive high order time schemes is to use a Taylor expansion in time and plug in the expression of the successive time derivatives obtained from the differential system resulting from the semi-discretization in space. Consider for example that after semi-discretization in space using Finite Differences (or any other space discretisation method) we obtain the differential systems

$$\frac{dU}{dt} = \mathcal{A}U, \text{ with } U = \begin{pmatrix} u_0(t) \\ \vdots \\ u_{n-1}(t) \end{pmatrix},$$

and $\mathcal{A}$ the appropriate matrix coming from the semi-discretization in space. Then a Taylor expansion in time up to order p yields

$$U(t_{n+1}) = U(t_n) + \Delta t \frac{dU}{dt}(t_n) + \cdots + \frac{\Delta t^p}{p!}\frac{d^pU}{dt^p}(t_n) + O(\Delta t^{p+1}).$$

Now if $\mathcal{A}$ does not depend on time and $\frac{dU}{dt} = \mathcal{A}U$, we get that

$$\frac{d^pU}{dt^p} = \mathcal{A}^pU, \text{ for any integer } p.$$

Hence, denoting U^n an approximation of $U(t_n)$, we get a time scheme of order p using the formula

$$U^{n+1} = U^n + \Delta t\mathcal{A}U^n + \cdots + \frac{\Delta t^p}{p!}\mathcal{A}^pU^n = (I + \Delta t\mathcal{A} + \cdots + \frac{\Delta t^p}{p!}\mathcal{A}^p)U^n. \tag{2.13}$$

For $p = 1$ this boils down to the standard explicit Euler scheme.

Writing U^n the solution in vector form at time t_n, we define the propagation matrix A such that

$$U^{n+1} = AU^n.$$

Proposition 7 *The numerical scheme defined by the propagation matrix A is stable if there exists $\tau > 0$ such that for all $\Delta t < \tau$ all eigenvalues of A are of modulus less or equal to 1.*

Stability of Taylor schemes. For a Taylor scheme of order p applied to $\frac{dU}{dt} = \mathcal{A}U$, we have $A = I + \Delta t\mathcal{A} + \cdots + \frac{\Delta t}{p!}\mathcal{A}^p$. Then denoting by λ an eigenvalue of $\mathcal{A}$, the corresponding eigenvalue of A is $\mu = 1 + \lambda\Delta t + \cdots + \lambda^p\frac{\Delta t^p}{p!}$. And one can plot the region of the complex plane in which $|\mu(\lambda\Delta t)| \leq 1$ using for example `ezplot` in Matlab, which are the stability regions. This means that the time scheme associate to the semi-discrete form $\frac{dU}{dt} = \mathcal{A}U$ is stable provided all the eigenvalues λ of $\mathcal{A}$ are such that $\lambda\Delta t$ is in the stability region.

Examples.

1. The Upwind scheme: $\frac{du_i(t)}{dt} = -a\frac{u_i(t)-u_{i-1}(t)}{\Delta x}$ corresponds to the circulant matrix $\mathcal{A}$ with $c_0 = -\frac{a}{\Delta x} = -c_1$. So its eigenvalues verify $\lambda_k\Delta t = -\frac{a\Delta t}{\Delta x}(1 - e^{\frac{2i\pi k}{n}})$.

Obviously, for any integer value of k, $\lambda_k \Delta t$ is on a circle in the complex plane of radius $\frac{a\Delta t}{\Delta x}$ centred at $(-\frac{a\Delta t}{\Delta x}, 0)$. The stability region of the explicit Euler method is the circle of radius 1 centred at $(-1, 0)$, so that in this case we see again that the scheme is stable provided $\frac{a\Delta t}{\Delta x} \leq 1$. For the higher order schemes the limit of the stability region is reached when the circle of the eigenvalues of $\mathcal{A}$ is tangent to the left side of the stability region. The radius corresponding to the maximal stability can thus be found by computing the second real root (in addition to 0) α of the equation $|\mu(\lambda\Delta t)| = 1$, see Fig. 2.2. We find that for the order 2 scheme $\alpha = -2$, so that the stability condition is the same as for the order 1 scheme. For the order 3 scheme we find that $\alpha = -2.5127$ and for the order 4 scheme we find that $\alpha = -2.7853$. The value of α corresponds to the diameter of the largest circle of eigenvalues that is still completely enclosed in the stability region. This yields the stability condition $\frac{a\Delta t}{\Delta x} \leq \frac{|\alpha|}{2}$. We notice that the maximal stable time step is larger for the schemes of order 3 and 4.

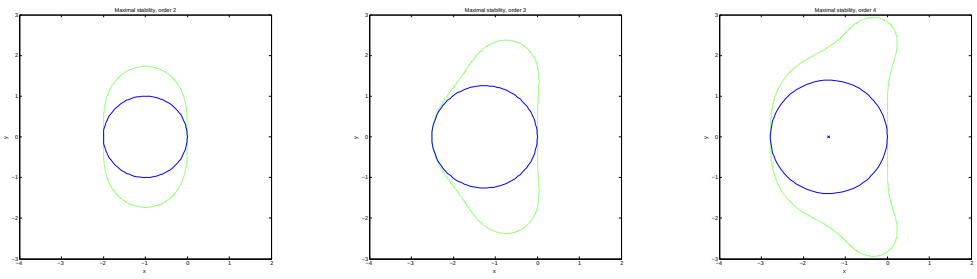

Figure 2.2: Location of eigenvalues (blue circle) corresponding to maximal stability zone for explicit time schemes of order 2, 3, 4 (left to right).

2. The centred scheme: $\frac{du_i(t)}{dt} = -a\frac{u_{i+1}(t)-u_{i-1}(t)}{2\Delta x}$ corresponds to the circulant matrix $\mathcal{A}$ with $c_1 = -\frac{a}{2\Delta x} = -c_{n-1}$. The corresponding eigenvalues are such that

$$\lambda_k \Delta t = -\frac{a\Delta t}{2\Delta x}(e^{\frac{2i\pi jk}{n}} - e^{\frac{-2i\pi jk}{n}}) = -\frac{ia\Delta t}{\Delta x}\sin\frac{2\pi jk}{n}.$$

Hence the eigenvalues are all purely imaginary and the modulus of the largest one is $\frac{a\Delta t}{\Delta x}$. The stability zones for the schemes of order 1 to 6 are represented in Fig. 2.3. Note that for the order 1 and 2 scheme the intersection of the stability zone with the imaginary axis is reduced to the point 0. So that when all eigenvalues are purely imaginary as is the case here, these schemes are not stable for any positive Δt. On the other hand the schemes of order 3 and 4 have a non vanishing stability zone on the imaginary axis, larger for the order 4 scheme. By computing the intersection of the curve $|\mu(\lambda\Delta t)| = 1$ with the imaginary axis we find the stability conditions for the order 3 scheme: $\frac{a\Delta t}{\Delta x} \leq \sqrt{3}$ and for the order 4 scheme $\frac{a\Delta t}{\Delta x} \leq 2\sqrt{2}$.

Remark 2 *The order 5 and 6 schemes are more problematic for eigenvalues of $\mathcal{A}$ on the imaginary axis as the zooms of Figure 2.4 tell us. Even though there is a part of the imaginary axis in the stability zone, there is also a part in the neighborhood of 0 which is not. Therefore small eigenvalues of A will lead to instability on longer time scales. This is problematic, as unlike usual Courant condition instability problems which reveal themselves very fast, this leads to a small growth in time.*

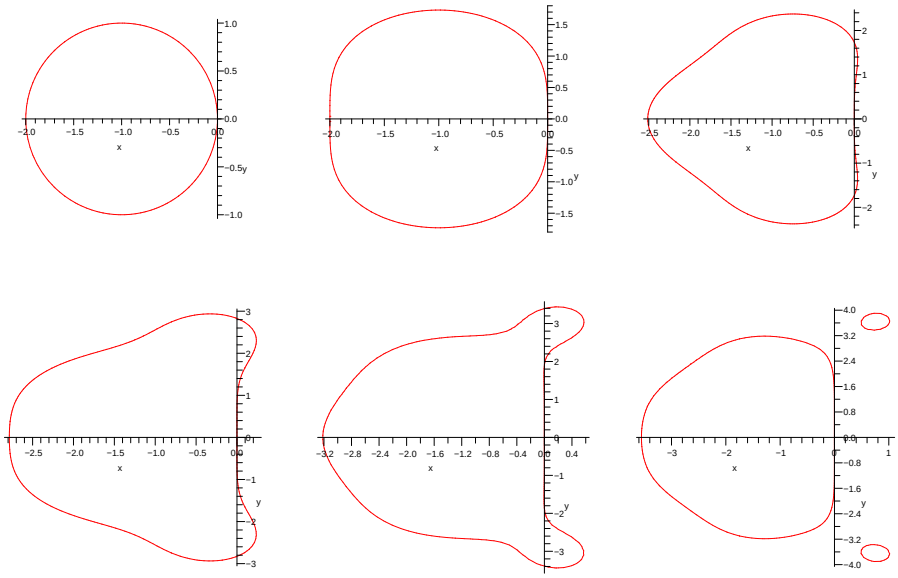

Figure 2.3: Stability zone for Taylor schemes. From top to bottom and left to right order 1, 2, 3, 4, 5, 6.

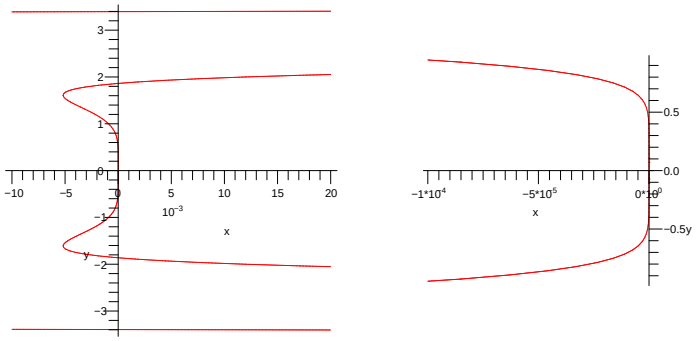

Figure 2.4: Stability zone for Taylor schemes. Zoom around imaginary axis. Left order 5, right order 6.

2.2 The Finite Element method

2.2.1 Theoretical background

The finite element method is based on a variational formulation, which consists in seeking the solution of a boundary value problem in some function space, typically H^1 or H^1_0 for the Poisson equation with Neumann or Dirichlet boundary conditions. This variational problem is reduced to a finite dimensional problem, which are the only ones that can be handled by a computer, by seeking the solution in a finite dimensional subspace of the original function space. The variational formulation itself being unmodified or slightly perturbed by a term with tends to 0 at convergence. For this reasons the mathematical tools for proving convergence are closely related to the tools for proving existence and uniqueness of the solution of the initial problem. A detailed description can be found for example in the book by Ern and Guermond that we follow.

Consider the variational problem: *Find $u \in W$ such that*

$$a(u, v) = f(v) \quad \forall v \in V. \tag{2.14}$$

For elliptic problems, with $V = W$, the classical theorem for this is the Lax-Milgram theorem that reads as follows.

Theorem 2 (Lax-Milgram) *Let V be a Hilbert space. Assume a is a continuous bilinear form on V and f is a continuous linear form on V and that a is coercive i.e.*

$$\exists \alpha > 0, \quad \forall u \in V, \quad a(u, u) \geq \alpha \|u\|_V^2.$$

Then the variational problem admits a unique solution and we have the a priori estimate

$$\forall f \in V', \quad \|u\|_V \leq \frac{1}{\alpha} \|f\|_{V'}.$$

In the case of hyperbolic problem it is often mandatory or more efficient to have the trial space, where the solution is sought, be different from the test space in which the test functions live. In this case the appropriate theoretical tool, called Banach-Nečas-Babuška (BNB) theorem in Ern and Guermond.

Theorem 3 (Banach-Nečas-Babuška) *Let W be a Banach space and let V be a reflexive Banach space. Assume a is a continuous bilinear form on $W \times V$ and f is a continuous linear form on V and that the following two hypotheses are verified*

1) $\exists \alpha > 0, \quad \inf_{w \in W} \sup_{v \in V} \dfrac{a(w, v)}{\|w\|_W \|v\|_V} \geq \alpha.$

2) $a(w, v) = 0 \quad \forall w \in W \Rightarrow v = 0.$

Then the variational problem admits a unique solution and we have the a priori estimate

$$\forall f \in V', \quad \|u\|_V \leq \frac{1}{\alpha} \|f\|_{V'}.$$

The Lax-Milgran theorem is a special case of the BNB theorem. Indeed if $V = W$ and a is coercive, then for any $w \in V \setminus \{0\}$ we have

$$\alpha \|w\|_V \leq \frac{a(w, w)}{\|w\|_V} \leq \sup_{v \in V} \frac{a(w, v)}{\|w\|_V}.$$

Dividing by $\|w\|_V$ and taking the infimum give condition 1 of BNB. Then if the condition on the left hand side of the implication of 2 is satisfied, we have for $w = v$ $a(v, v) = 0$ and because of coercivity this implies $v = 0$.

Condition 1 will play an essential role throughout the lecture. This condition being satisfied at the discrete level with a constant α that does not depend on the mesh size being essential for a well behaved Finite Element method. This condition is usually called the inf-sup condition in the literature and this is the name we will use in this lecture. It can be written equivalently

$$\alpha\|w\|_W \leq \sup_{v\in V} \frac{a(w,v)}{\|v\|_V} \quad \forall w \in W. \tag{2.15}$$

And often, a simple way to verify it is, given any $w \in W$, to find a specific $v(w)$ depending on w such that

$$\alpha\|w\|_W \leq \frac{a(w, v(w))}{\|v(w)\|_V} \leq \sup_{v\in V} \frac{a(w,v)}{\|v\|_V}$$

with a constant α independent of w. For example, when a is coercive, with coercivity constant α, the inf-sup condition is proven by taking $v(w) = w$.

Let us now come to the Galerkin discretisation. The principle is simply to construct finite dimensional subspaces $W_h \subset W$ and $V_h \subset V$ and to write the variational formulation (2.14) replacing W by W_h and V by V_h. The finite dimensional space W_h is usually called *trial space* and V_h is called *test space*. Expanding the functions on bases of W_h and V_h this yields a finite dimensional linear system that can be solved with standard methods if a is bilinear and f is linear. The method can also be extended to the nonlinear case.

Let us denote by u the solution of the variational problem in W: $a(u, v) = f(v) \quad \forall v \in V$, and u_h the solution of the variational problem in W_h: $a(u_h, v_h) = f(v_h) \quad \forall v \in V_h$. Then we get by linearity and because $V_h \subset V$ that

$$a(u - u_h, v_h) = 0, \quad \forall v_h \in V_h.$$

This condition is called Galerkin orthogonality.

We are now ready to prove using simple arguments a convergence theorem for the Galerkin approximation:

Theorem 4 *Let W, $W_h \subset W$, V and $V_h \subset V$ be Banach spaces and let a be a continuous bilinear form on $W \times V$ with continuity constant $C > 0$. Assume the exact solution $u \in W$ and the approximate solution $u_h \in W_h$ satisfy the Galerkin orthogonality condition:*

$$a(u - u_h, v_h) = 0, \quad \forall v_h \in V_h,$$

and that a satisfies the discrete inf-sup condition for $\alpha > 0$

$$\alpha\|w_h\|_W \leq \sup_{v_h\in W_h} \frac{a(w_h, v_h)}{\|v_h\|_V} \quad \forall w_h \in W_h.$$

Then the following error estimate holds:

$$\|u - u_h\|_W \leq (1 + \frac{C}{\alpha}) \inf_{w_h\in W_h} \|u - w_h\|_W.$$

Proof. For any $w_h \in W_h$ we have

$$\|u - u_h\|_W \leq \|u - w_h\|_W + \|w_h - u_h\|_W.$$

Then, using first the discrete inf-sup condition and then the Galerkin orthogonality

$$\|w_h - u_h\|_W \leq \frac{1}{\alpha} \sup_{v_h \in W_h} \frac{a(w_h - u_h, v_h)}{\|v_h\|_V} = \frac{1}{\alpha} \sup_{v_h \in W_h} \frac{a(w_h - u, v_h)}{\|v_h\|_V}.$$

Finally the continuity of the bilinear form a can be expressed as

$$a(w_h - u, v_h) \leq C\|w_h - u\|_W \|v_h\|_V.$$

It then follows that

$$\|u - u_h\|_W \leq (1 + \frac{C}{\alpha})\|u - w_h\|_W.$$

This being true for all $w_h \in W_h$ the result follows by taking the infimum. ∎

The problem is now reduced to finding a trial space W_h which approximates well W. This is typically done in the finite element method by choosing an approximation space containing piecewise polynomials of degree k in which case it can be proven that $\inf_{w_h \in W_h} \|u - w_h\|_{L^2} \leq ch^{k+1}$, where h is related to the cell size. Details can be found in any Finite Element book, like for example Ern and Guermond. If the discrete inf-sup constant α does not depend on h, the error in the Finite Element approximation is the same, up to a constant, as the best approximation. The error is said to be optimal in this case.

Remark 3 *In some situations the bilinear form a and the linear form f need to be approximated in the finite dimensional context, but we will not consider these here. In practice if the approximations are consistent the theory remains similar.*

2.2.2 Galerkin discretisation of the 1D advection-diffusion equation

The Galerkin discretisation is based on a weak (or variational form of the equation). In order to obtain the weak form, the idea is to multiply by a smooth test function and integrate over the whole domain, with a possible integration by parts to eliminate the highest derivatives. As in the case of Finite Differences, we consider here only a semi-discretisation in space by Finite Elements. The discretisation in time being handled by an appropriate ODE solver.

Let us describe it on the advection-diffusion problem (assuming periodic boundary conditions on the domain $[0, L]$):

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0.$$

Multiplying by a test function v which does not depend on t and integrating, with an integration by parts in the last integral and periodic boundary conditions, yields

$$\frac{d}{dt}\int_0^L uv \,\mathrm{d}x + a\int_0^L \frac{\partial u}{\partial x} v \,\mathrm{d}x + \nu\int_0^L \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} \,\mathrm{d}x = 0.$$

The natural space in which this formulation is defined is $H^1([0,L]) = \{u \in L^2([0,L]) \mid \frac{\partial u}{\partial x} \in L^2([0,L])\}$. The variational problem thus reads *Find* $u \in H^1([0,L])$ *such that*

$$\frac{d}{dt}\int_0^L uv\,\mathrm{d}x + a\int_0^L \frac{\partial u}{\partial x}v\,\mathrm{d}x + \nu\int_0^L \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,\mathrm{d}x = 0 \quad \forall v \in H^1([0,L]).$$

Now in order to define a Finite Element approximation, we need to construct a sequence of subspaces V_h of $H^1([0,L])$ which is dense in $H^1([0,L])$ (in order to get convergence). One subspace is of course enough to compute a numerical approximation. There are many ways to construct such subspaces. The classical Lagrange Finite Element method consists in building a mesh of the computational domain and to assume that the approximating function is polynomial say of degree k in each cell. For the piecewise polynomial function space to be a included in $H^1([0,L])$ the only additional requirement is that the functions are continuous at the cell interfaces. So the subspace V_h on the mesh $x_0 < x_1 < \cdots < x_{n-1}$ is defined by

$$V_h = \left\{v_h \in C^0([a,b]) \mid v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_k([x_i,x_{i+1}])\right\}.$$

In order to express a function $v_h \in V_h$ we need a basis of V_h. This can be constructed easily combining bases of $\mathbb{P}_k([x_i,x_{i+1}])$. In order to impose the continuity requirement at the cell interface, the simplest is to use a Lagrange basis $\mathbb{P}_k([x_i,x_{i+1}])$ with interpolation points on the edge of the intervals. Given $k+1$ interpolation points $x_i = y_0 < y_1 < \cdots < y_k = x_{i+1}$ the Lagrange basis functions of degree k denoted by $l_{k,i}$, $0 \le i \le k$, are the unique polynomials of degree k verifying $l_{k,i}(y_j) = \delta_{i,j}$. Because of this property, any polynomial $p(x) \in \mathbb{P}_k([x_i,x_{i+1}])$ can be expressed as $p(x) = \sum_{j=0}^k p(y_j)l_{j,k}(x)$ and conversely any polynomial $p(x) \in \mathbb{P}_k([x_i,x_{i+1}])$ is uniquely determined by its values at the interpolation points y_j, $0 \le j \le k$. Hence in order to ensure the continuity of the piecewise polynomial at the cell interface x_i it is enough that the values of the polynomials on both sides of x_i have the same value at x_i. This constraint removes one degree of freedom in each cell, so that the total dimension of V_h is nk and the functions of V_h are uniquely defined in each cell by their value at the degrees of freedom (which are the interpolation points) in all the cells. The basis functions denoted of V_h denoted by $(\varphi_i)_{0\le j\le nk}$ are such that their restriction on each cell is a Lagrange basis function.

Note that for $k = 1$, corresponding to $\mathbb{P}_1$ finite elements, the degrees of freedom are just the grid points. For higher order finite elements internal degrees of freedom are needed. For stability and conveniency issues this are most commonly taken to be the Gauss-Lobatto points on each cell.

In order to obtain a discrete problem that can be solved on a computer, the Galerkin procedure consist in replacing H^1 by V_h in the variational formulation. The discrete variational problem thus reads: *Find* $u_h \in V_h$ *such that*

$$\frac{d}{dt}\int_0^L uv\,\mathrm{d}x + a\int_0^L \frac{\partial u}{\partial x}v\,\mathrm{d}x + \nu\int_0^L \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,\mathrm{d}x = 0 \quad \forall v_h \in V_h.$$

Now expressing u_h (and v_h) in the basis of V_h as $u_h(t,x) = \sum_{j=1}^{nk} u_j(t)\varphi_j(x)$, $v_h(x) = \sum_{j=1}^{nk} v_j\varphi_j(x)$ and plugging these expression in the variational formulation, denoting by

$U = (u_0, u_1, \cdots, u_{nk})^T$ and similarly for V yields: *Find* $U \in \mathbb{R}^{nk}$ *such that*

$$\frac{d}{dt}\sum_{i,j} u_j v_i \int_0^L \varphi_i(x)\varphi_j(x)\,\mathrm{d}x + a\sum_{i,j} u_j v_i \int_0^L \frac{\partial\varphi_j(x)}{\partial x}\varphi_i(x)\,\mathrm{d}x \\ + \nu\sum_{i,j} u_j v_i \int_0^L \frac{\partial\varphi_i(x)}{\partial x}\frac{\partial\varphi_j(x)}{\partial x}\,\mathrm{d}x = 0 \quad \forall V \in \mathbb{R}^{nk},$$

which can be expressed in matrix form

$$V^T(M\frac{dU}{dt} + DU + AU) = 0 \quad \forall V \in \mathbb{R}^{nk},$$

which is equivalent to

$$M\frac{dU}{dt} + DU + AU = 0$$

where the square $nk \times nk$ matrices are defined by

$$M = (\int_0^L \varphi_j(x)\varphi_i(x)\,\mathrm{d}x)_{i,j}, \quad D = (a\int_0^L \frac{\partial\varphi_j(x)}{\partial x}\varphi_i(x)\,\mathrm{d}x)_{i,j}, \quad A = (\nu\int_0^L \frac{\partial\varphi_i(x)}{\partial x}\frac{\partial\varphi_j(x)}{\partial x}\,\mathrm{d}x)_{i,j}.$$

Note that these matrices can be computed exactly as they involve integration of polynomials on each cell. Moreover because the Gauss-Lobatto quadrature rule is exact for polynomials of degree up to $2k-1$, both A and D can be computed exactly with the Gauss-Lobatto quadrature rule. Moreover, approximating the mass matrix M with the Gauss-Lobatto rule introduces an error which does not decrease the order of accuracy of the scheme [4] and has the big advantage of yielding a diagonal matrix. This is what is mostly done in practice.

Matrix Assembly. Usually for Finite Elements the matrices M, D and A are computed from the corresponding elementary matrices which are obtained by change of variables onto the reference element $[-1, 1]$ for each cell. So

$$\int_0^L \varphi_i(x)\varphi_j(x)\,\mathrm{d}x = \sum_{\nu=0}^{n-1}\int_{x_\nu}^{x_{\nu+1}} \varphi_i(x)\varphi_j(x)\,\mathrm{d}x,$$

and doing the change of variable $x = \frac{x_{\nu+1}-x_\nu}{2}\hat{x} + \frac{x_{\nu+1}+x_\nu}{2}$, we get

$$\int_{x_\nu}^{x_{\nu+1}} \varphi_i(x)\varphi_j(x)\,\mathrm{d}x = \frac{x_{\nu+1}-x_\nu}{2}\int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})\,\mathrm{d}\hat{x},$$

where $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{\nu+1}-x_\nu}{2}\hat{x} + \frac{x_{\nu+1}+x_\nu}{2})$. The local indices α on the reference element go from 0 to k and the global numbers of the basis functions not vanishing on element ν are $j = k\nu + \alpha$. The $\hat{\varphi}_\alpha$ are the Lagrange polynomials at the Gauss-Lobatto points in the interval $[-1, 1]$.

The mass matrix in V_h can be approximated with no loss of order of the finite element approximation using the Gauss-Lobatto quadrature rule. Then because the products $\hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})$ vanish for $\alpha \neq \beta$ at the Gauss-Lobatto points by definition of the

$\hat{\varphi}_\alpha$ which are the Lagrange basis functions at these points, the elementary matrix M is diagonal and we have

$$\int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})^2 \,\mathrm{d}\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL} \varphi_\alpha(\hat{x}_\beta)^2 = w_\alpha^{GL}$$

using the quadrature rule, where w_α^{GL} is the Gauss-Lobatto weight at Gauss-Lobatto point $(\hat{x}_\alpha) \in [-1, 1]$. So that finally $\hat{M} = diag(w_0^{GL}, \dots w_k^{GL})$ is the matrix with $k+1$ lines and columns with the Gauss-Lobatto weights on the diagonal.

Let us now compute the elements of D. As previously we go back to the interval $[-1, 1]$ with the change of variables $x = \frac{x_{\nu+1}-x_\nu}{2}\hat{x} + \frac{x_{\nu+1}+x_\nu}{2}$ and we define $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{\nu+1}-x_\nu}{2}\hat{x} + \frac{x_{\nu+1}+x_\nu}{2})$. Note that a global basis function φ_i associated to a grid point has a support which overlaps two cells and is associated to two local basis functions. Thus one needs to be careful to add the two contributions as needed in the final matrix.

We get $\hat{\varphi}'_\alpha(\hat{x}) = \frac{x_{\nu+1}-x_\nu}{2}\varphi'_i(\frac{x_{\nu+1}-x_\nu}{2}(\hat{x}+1) + x_\nu)$. It follows that

$$\int_{x_\nu}^{x_{\nu+1}} \varphi'_j(x)\varphi_i(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{2}{x_{\nu+1}-x_\nu} \hat{\varphi}'_\beta(\hat{x})\hat{\varphi}_\alpha(\hat{x}) \frac{x_{\nu+1}-x_\nu}{2}\,\mathrm{d}\hat{x} = \int_{-1}^{1} \hat{\varphi}'_\beta(\hat{x})\hat{\varphi}_\alpha(\hat{x})\,\mathrm{d}\hat{x}.$$

The polynomial $\hat{\varphi}_\alpha(\hat{x})$ is of degree k so that $\hat{\varphi}'_\beta(\hat{x})$ is of degree $k-1$ so that the Gauss-Lobatto quadrature rule with $k+1$ points is exact for the product which is of order $2k-1$. Using this rule

$$\int_{-1}^{1} \hat{\varphi}'_\beta(\hat{x})\hat{\varphi}_\alpha(\hat{x})\,\mathrm{d}\hat{x} = \sum_{m=0}^{k} w_m^{GL} \hat{\varphi}'_\beta(\hat{x}_m)\hat{\varphi}_\alpha(\hat{x}_m) = w_\alpha^{GL}\hat{\varphi}'_\beta(\hat{x}_\alpha),$$

As before, because $\hat{\varphi}_\alpha$ are the Lagrange polynomials at the Gauss-Lobatto points, only the value at x_α in the sum is one and the others are 0. On the other hand evaluating the derivatives of the Lagrange polynomial at the Gauss-Lobatto points at these Gauss-Lobatto points can be done using the formula

$$\hat{\varphi}'_\alpha(\hat{x}_\beta) = \frac{p_\beta/p_\alpha}{\hat{x}_\beta - \hat{x}_\alpha} \text{ for } \beta \neq \alpha \text{ and } \hat{\varphi}'_\alpha(\hat{x}_\alpha) = -\sum_{\beta\neq\alpha} \hat{\varphi}'_\beta(\hat{x}_\alpha),$$

where $p_\alpha = \prod_{\beta\neq\alpha}(\hat{x}_\alpha - \hat{x}_\beta)$. This formula is obtained straightforwardly by taking the derivative of the explicit formula for the Lagrange polynomial

$$\hat{\varphi}_\alpha(\hat{x}) = \frac{\displaystyle\prod_{\beta\neq\alpha}(\hat{x} - \hat{x}_\beta)}{\displaystyle\prod_{\beta\neq\alpha}(\hat{x}_\alpha - \hat{x}_\beta)}$$

and using this expression at the Gauss-Lobatto point $\hat{x}_\beta \neq \hat{x}_\alpha$. We refer to [3] for a detailed description.

We can now conclude with the computation of the stiffness matrix A. Having already computed the expression of the change of variable of the derivatives we can quickly go to the result. We have in each element

$$\begin{aligned}\int_{x_\nu}^{x_{\nu+1}} \varphi'_j(x)\varphi'_i(x)\,\mathrm{d}x &= \int_{-1}^{1} \left(\frac{2}{x_{\nu+1}-x_\nu}\right)^2 \hat{\varphi}'_\beta(\hat{x})\hat{\varphi}'_\alpha(\hat{x}) \frac{x_{\nu+1}-x_\nu}{2}\,\mathrm{d}\hat{x} \\ &= \frac{2}{x_{\nu+1}-x_\nu}\int_{-1}^{1} \hat{\varphi}'_\beta(\hat{x})\hat{\varphi}'_\alpha(\hat{x})\,\mathrm{d}\hat{x} = \frac{2}{x_{\nu+1}-x_\nu}\sum_{m=0}^{k} w_m^{GL}\hat{\varphi}'_\beta(\hat{x}_m)\hat{\varphi}'_\alpha(\hat{x}_m).\end{aligned}$$

As the polynomial being integrated is of degree $2(k-1) = 2k-2$ the Gauss-Lobatto quadrature is exact. Here no simplifications occurs in the sum, which has to be computed. Still the expressions of the derivatives at the Gauss-Lobatto points computed above can then be plugged in.

Time advance and stability. At the end we get as for the finite difference method a system of differential equations that can be solved with any ODE solver. Let us make a few remarks concerning the stability of the scheme (once discretised in time). As we saw previously this depends on whether the eigenvalues of the matrix $\mathcal{A}$ is included in the stability zone of the ODE solver. Here $\mathcal{A} = -M^{-1}(D+A)$. Note that the matrices M and A are obviously symmetric and thus have only real eigenvalues. On the other hand, for periodic boundary conditions and integration by parts, yields that $\int \frac{\partial \varphi_j(x)}{\partial x}\varphi_i(x)\,\mathrm{d}x = -\int \varphi_j(x)\frac{\partial \varphi_i(x)}{\partial x}\,\mathrm{d}x$. Hence D is skew symmetric and has only imaginary eigenvalues. Remember that the stability zones of our explicit ODE solvers lie mostly on the left-hand side of the imaginary axis in the complex plane, and that only the third and fourth order schemes have a stability zone including a part of the imaginary axis. Hence in the pure advection case $\nu = 0$ $\mathcal{A}$ has purely imaginary eigenvalues and the order one and two time schemes are always unstable. In order to stabilise the method a procedure that is often used with a finite element discretisation of a pure advection problem is to add a diffusive term that goes to 0 with the cell size, i.e take $\nu = \alpha\Delta x$, in this case a small negative real part is added to the eigenvalues which are thus pushed into the left half of the complex plane and the stability zone is enhanced.

Suboptimality of Finite Element approximation of advection. In the case when $W_h = V_h$ are standard Lagrange Finite Elements the discrete inf-sup constant α_h is of the order of the space step h. Then one order of approximation is lost, the approximation is not optimal. For $\mathbb{P}_1$ finite elements Ern and Guermond [7] (Theorem 5.3. p 222) prove that

$$c_1 h < \inf_{u_h \in V_h} \sup_{v_h \in V_h} \frac{a(u_h, v_h)}{\|u_h\|_1 \|v_h\|_1} < c_2 h,$$

where c_1 and c_2 are two constants independent of h and $\|.\|_1$ denotes the H^1 norm.

Better approximations can be found by changing the test space while keeping the same trial space, in order to find a discrete inf-sup constant that does not depend on h.

2.3 The Discontinuous Galerkin (DG) method

The DG method represents the unknowns like the Finite Element method by piecewise polynomial functions, but unlike Finite Element methods the polynomials are discontinuous at the cell interfaces and a numerical flux is defined at the cell interface in the same way as for Finite Volume methods.

So on each cell the discrete unknown u_h is represented as a linear combination of well chosen basis functions of the space of polynomials of degree k $\mathbb{P}_k$. The dimension of this space is $k+1$. As no continuity is enforced at the element interface, there is no constraint on the basis functions and generally two kinds of basis functions are used: either the Legendre polynomials which form an orthogonal basis, we then speak of *modal DG* or one can use a Lagrange basis defined on a set of interpolation points within the cell we then speak of *nodal DG*. The interpolation points are generally chosen to be either the

Gauss points or the Gauss-Lobatto points which are both very convenient as they allow to express the integrals appearing in the formulation exactly (for Gauss) or almost (for Gauss-Lobatto) using the associated quadrature rule.

For the derivation of a DG scheme, the equation is multiplied by a polynomial test function on each cell and and integration by parts is used so that a boundary term appears which will allow the coupling between two neighbouring cells. Let us apply it here to the conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0.$$

We then get

$$\begin{aligned} &\frac{d}{dt}\int_{x_\nu}^{x_{\nu+1}} uv \,\mathrm{d}x + \int_{x_\nu}^{x_{\nu+1}} \frac{\partial f(u)}{\partial x} v \,\mathrm{d}x \\ &= \frac{d}{dt}\int_{x_\nu}^{x_{\nu+1}} uv \,\mathrm{d}x - \int_{x_\nu}^{x_{\nu+1}} f(u)\frac{\partial v}{\partial x} \,\mathrm{d}x + (f(u(x_{\nu+1}))v(x_{\nu+1}) - f(u(x_\nu))v(x_\nu)) = 0. \end{aligned} \tag{2.16}$$

The DG method has then two key ingredients.

1. Choose on each cell a finite dimensional representation, usually by a polynomial as said previously.

2. Define a unique numerical flux denoted by $g_\nu = g(u_L(x_\nu), u_R(x_\nu)))$ at the cell interface which is constructed from the two values coming from the cells sharing the interface on the left and on the right. Indeed the approximation of u being discontinuous at the cell interface, the values $u(x_\nu)$ and $f(u(x_\nu))$ are not defined in a natural way and ingredient of the scheme is to approximate the flux $f(u(x_\nu))$ by the numerical flux g_ν

At the interface between two cells x_ν, the Discontinuous Galerkin approximation provides two values of $u(x_\nu)$, u_L coming from the approximation of u on the left of x_ν and u_R corresponding to the value at x_ν from the right-hand cell. The numerical flux at each cell interface g_ν needs to be consistent with $f(x_\nu)$, i.e. $g_\nu = f(u(x_\nu)) + O(\Delta x^p)$ for some positive integer p. A numerical flux of order 2 is the centred flux $g_\nu = \frac{1}{2}(f(u_L) + f(u_R))$. The centred flux amounts to projecting the discontinuous approximation to a continuous Finite Element basis and with yield a skew symmetric derivative matrix. Thus this scheme is unstable for explicit time discretisations of order 1 and 2. In order to get stable scheme in this case, we need to introduce the notion of unwinding like for Finite Differences. This can be done very easily in the definition of the numerical flux by simply choosing the value of u in the upwind cell only to define the numerical flux. We have $\frac{\partial f(u)}{\partial x} = f'(u)\frac{\partial u}{\partial x}$. This means that locally at each cell interface the direction of the transport is defined by the sign of $f'(u)$ (in the case of the linear advection $f'(u) = a$ and the upwind direction is determined by the sign of a). So the upwind numerical flux is defined by

$$g_\nu = g(u_L(x_\nu), u_R(x_\nu))) = \left| \begin{array}{lll} f(u_L) & \text{if} & f'(\frac{u_L+u_R}{2}) \geq 0, \\ f(u_R) & \text{if} & f'(\frac{u_L+u_R}{2}) < 0. \end{array} \right.$$

Choosing as local representation for u and the test function v the Lagrange polynomials at the Gauss-Lobatto points simplifies the computation of the fluxes, as in this case

only the Lagrange polynomial associated to the edge node does not vanish at the edge. This situation is different when using Legendre polynomials or Lagrange polynomials at only interior nodes (like the Gauss points). Note however that Legendre polynomials have the advantage of having exactly a diagonal mass matrix. This is obtained also with Lagrange polynomials at the Gauss-Lobatto points but in this case at the price of a small quadrature error.

As opposite to the Finite Element method, only local matrices on each element, in practice only the elementary matrices on the $[-1,1]$ reference interval need to be assembled. The elementary mass matrix $\hat{M}$ on cell on the reference interval has the components

$$\hat{M}_{\alpha,\beta} = \int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})\,\mathrm{d}x, \qquad 0 \le \alpha, \beta \le k.$$

When the basis functions are the Legendre polynomials which form an orthonormal basis.

The mass matrix in V_h can be approximated with no loss of order of the finite element approximation using the Gauss-Lobatto quadrature rule. Then because the products $\hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})$ vanish for $\alpha \neq \beta$ at the Gauss-Lobatto points by definition of the $\hat{\varphi}_\alpha$ which are the Lagrange basis functions at these points, the elementary matrix M is diagonal and we have

$$\int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})^2\,\mathrm{d}\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL}\varphi_\alpha(\hat{x}_\beta)^2 = w_\alpha^{GL}$$

using the quadrature rule, where w_α^{GL} is the Gauss-Lobatto weight at Gauss-Lobatto point $(\hat{x}_\alpha) \in [-1,1]$. So that finally $\hat{M} = diag(w_0^{GL}, \dots w_k^{GL})$ is the matrix with $k+1$ lines and columns with the Gauss-Lobatto weights on the diagonal. From this matrix, the local mass matrix $M_{\nu+\frac{1}{2}}$ on cell $[x_\nu, x_{\nu+1}]$ can be expressed as

$$M_{\nu+\frac{1}{2}} = \frac{x_{\nu+1} - x_\nu}{2}\hat{M}.$$

Let us denote by $K_{\nu+\frac{1}{2}}$ the local matrix containing the derivative of v. In order to get an expression for the components of $K_{\nu+\frac{1}{2}}$ we introduce the local basis functions and compute using again the affine change of variable to the reference interval $[-1,1]$:

$$\int_{x_\nu}^{x_{\nu+1}} f(u)\frac{\partial\varphi_i}{\partial x}\,\mathrm{d}x = \int_{-1}^{1} f(u(x))\frac{2}{x_{\nu+1}-x_\nu}\hat{\varphi}'_\alpha(\hat{x})\frac{x_{\nu+1}-x_\nu}{2}\,\mathrm{d}\hat{x} = \int_{-1}^{1} f(u(x))\hat{\varphi}'_\alpha(\hat{x})\,\mathrm{d}\hat{x}.$$

Then using the Gauss-Lobatto quadrature rule this becomes

$$\int_{-1}^{1} f(u(\hat{x}))\hat{\varphi}'_\alpha(\hat{x})\,\mathrm{d}\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL} f(u(\hat{x}_\beta))\hat{\varphi}'_\alpha(\hat{x}_\beta) = \sum_{\beta=0}^{k} w_\beta^{GL} f(u_\beta)\hat{\varphi}'_\alpha(\hat{x}_\beta),$$

where $u_\beta = u(\hat{x}_\beta)$ is the βth component of u on the Lagrange basis. Denoting $U_{\nu+\frac{1}{2}} = (u_0, u_1, \dots, u_k)$ on the cell $[x_\nu, x_{\nu+1}]$, thus defining the component of matrix $K_{\nu+\frac{1}{2}}$ at line α and column β as being $w_\beta^{GL}\hat{\varphi}'_\alpha(\hat{x}_\beta)$, we get that

$$\int_{-1}^{1} f(u(\hat{x}))\hat{\varphi}'_\alpha(\hat{x})\,\mathrm{d}\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL} f(u_\beta)\hat{\varphi}'_\alpha(\hat{x}_\beta) = (K_{\nu+\frac{1}{2}} f(U_{\nu+\frac{1}{2}}))_\alpha,$$

where $f(U_{\nu+\frac{1}{2}}) = (f(u_0), f(u_1), \dots, f(u_k))$.

Remark 4 *Because the Gauss-Lobatto quadrature is exact in this case, we notice that the matrix* $K_{\nu+\frac{1}{2}}$ *is exactly the matrix associated to the compotents*

$$\int_{x_\nu}^{x_{\nu+1}} \varphi_j(x) \frac{\partial \varphi_i}{\partial x} \,\mathrm{d}x = \int_{-1}^{1} \varphi_\beta(\hat{x}) \hat{\varphi}'_\alpha(\hat{x}) \,\mathrm{d}\hat{x} = w_\beta^{GL} \hat{\varphi}'_\alpha(\hat{x}_\beta).$$

We also notice that this matrix does not depend on the specific interval and is equal to the matrix on the reference element $\hat{K} = K_{\nu+\frac{1}{2}}$ *for all* ν*.*

Now plugging all this into the formula (2.16) we get on each cell

$$V^T_{\nu+\frac{1}{2}} M_{\nu+\frac{1}{2}} \frac{dU_{\nu+\frac{1}{2}}}{dt} = V^T_{\nu+\frac{1}{2}} \hat{K} f(U_{\nu+\frac{1}{2}}) - (g_{\nu+1} v_k - g_\nu v_0).$$

Then introducing the vector $G_{\nu+\frac{1}{2}} \in \mathbb{R}^k$ whose only non zero components are the first which is g_ν and the last which is $g_{\nu+1}$, we get the following system of ODE

$$\frac{x_{\nu+1} - x_\nu}{2} \hat{M} \frac{dU_{\nu+\frac{1}{2}}}{dt} = \hat{K} f(U_{\nu+\frac{1}{2}}) + G_{\nu+\frac{1}{2}}.$$

The numerical flux g_ν depends on values of u coming from the neighbouring cell, this is where the coupling between the cells takes place. The matrix $\hat{M}$ being diagonal there is no linear system to solve. Simple examples of fluxes in the linear case $f(u) = au$ are the same as for the finite volume method with the centred or upwind fluxes, the two values being used here are the values of u on the interface coming from the two cells sharing the interface, this will be the local value of u_k from the left cell and the local value of u_0 from the right cell.

Chapter 3

Linear systems

3.1 Expressions of the Maxwell equations

3.1.1 The 3D Maxwell equations

The general expression for the Maxwell equations reads

$$
\begin{aligned}
-\frac{\partial \mathbf{D}}{\partial t} + \nabla \times \mathbf{H} &= \mathbf{J}, && (3.1)\\
\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} &= 0, && (3.2)\\
\nabla \cdot \mathbf{D} &= \rho, && (3.3)\\
\nabla \cdot \mathbf{B} &= 0, && (3.4)\\
\mathbf{D} &= \varepsilon \mathbf{E} && (3.5)\\
\mathbf{B} &= \mu \mathbf{H}. && (3.6)
\end{aligned}
$$

Initial and boundary conditions are needed in addition to fully determine the solution.

The last two relations are called the constitutive laws and permittivity ϵ and the permeability μ depend on the material. They can be discontinuous if several materials are considered. In vacuum $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$ are constants and they verify $\varepsilon_0 \mu_0 c^2 = 0$ where c is the speed of light. Then $\mathbf{D}$ and $\mathbf{H}$ are generally eliminated of the system.

Note that taking the divergence of (3.1) yields

$$\frac{\partial \nabla \cdot \mathbf{D}}{\partial t} = -\nabla \cdot \mathbf{J} = \frac{\partial \rho}{\partial t}$$

using the continuity equation $\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$. Hence if (3.3) is satisfied at time $t = 0$ it will be satisfied at all times. In the same way if $\nabla \cdot \mathbf{B} = 0$ at the initial time it will remain so for all times.

3.1.2 The 2D Maxwell equations

We consider the Maxwell equations in vacuum on a two dimensional domain Ω on which the fields are independent of the z variable. Then the electromagnetic field obeys to two sets of decoupled equations, the first of which involving the (E_x, E_y, B_z) components (TE mode) and the second involving the (E_z, B_x, B_y) components (TM mode). We present

only the first system, the other can be dealt with in a similar manner. This system reads

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\, B = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{3.7}$$

$$\frac{\partial B}{\partial t} + \operatorname{curl} \mathbf{E} = 0, \tag{3.8}$$

$$\operatorname{div} \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{3.9}$$

where $\mathbf{E} = (E_x, E_y)^T$, $B = B_z$, $\mathbf{curl}\, B_z = (\partial_y B_z, -\partial_x B_z)^T$, $\operatorname{curl} \mathbf{E} = \partial_x E_y - \partial_y E_x$, and $\operatorname{div} \mathbf{E} = \partial_x E_x + \partial_y E_y$.

In 2D Maxwell's equations can be split into two independent parts called the TE (transverse electric) mode and TM (transverse magnetic) mode. Mathematically they have the same structure. So it will be enough to study one of them. Let us for example consider the TE mode that writes

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\, B_z = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{3.10}$$

$$\frac{\partial B_z}{\partial t} + \operatorname{curl} \mathbf{E} = 0, \tag{3.11}$$

$$\operatorname{div} \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{3.12}$$

A condition for well-posedness of the Maxwell equations is that the sources $\mathbf{J}$ and ρ verify the continuity equation

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{J} = 0. \tag{3.13}$$

3.1.3 The 1D Maxwell equations

The system can be further decoupled in 1D, assuming that the fields only depend on x. Then (3.7) becomes

$$\frac{\partial E_x}{\partial t} = -\frac{1}{\varepsilon_0} J_x, \tag{3.14}$$

$$\frac{\partial E_y}{\partial t} + c^2 \frac{\partial B_z}{\partial x} = -\frac{1}{\varepsilon_0} J_y, \tag{3.15}$$

$$\frac{\partial B_z}{\partial t} + \frac{\partial E_y}{\partial x} = 0, \tag{3.16}$$

$$\frac{\partial E_x}{\partial x} = \frac{\rho}{\varepsilon_0}. \tag{3.17}$$

Note that here we decouple completely the propagative part of the electric field which is in 1D only E_y from its "static" part E_x. Components E_y and B_z are coupled by equations (3.15) and (3.16) and E_x is given either by the first component of the Ampère equation (3.14) or by Gauss's law (3.17), which are equivalent provided the initial condition satisfies Gauss's law and the 1D continuity equation $\frac{\partial \rho}{\partial t} + \frac{\partial J_x}{\partial x} = 0$, which are compatibility conditions.

3.1.4 Mixed Finite Element discretisation

We shall construct an arbitrary order mixed Finite Element approximation of the 1D Maxwell equations (3.15)-(3.16). For this we define a mesh $0 = x_0 < x_1 < x_2 < \cdots <$

$x_{N-1} < L$ of the periodic interval $[0, L[$. As we consider periodic boundary conditions the values at L will be the values at 0.

A Finite Element method is based on a variational formulation. In the case of mixed Finite Elements, one of the equation is kept in strong form (in our case (3.15)) and the other is put in weak form by integrating by parts after having multiplied by a test function. This yields the following variational formulation:
Find $(E, B) \in H^1_\sharp([0, L]) \times L^2_\sharp([0, L])$ *such that*

$$\frac{d}{dt}\int_0^L E(x,t)F(x)\,dx - c^2\int_0^L B(x,t)\frac{\partial F}{\partial x}(x)\,dx \tag{3.18}$$
$$= -\frac{1}{\varepsilon_0}\int_0^L J(x,t)F(x)\,dx \quad \forall F \in H^1_\sharp([0,L]),$$
$$\frac{d}{dt}\int_0^L B(x,t)C(x)\,dx + \int_0^L \frac{\partial E}{\partial x}(x,t)C(x)\,dx = 0 \quad \forall C \in L^2_\sharp([0,L]). \tag{3.19}$$

The $\sharp$ subscript stands for spaces of periodic functions.

We define the discrete subpaces $V_k \subset H^1_\sharp([0, L])$ and $W_k \subset L^2_\sharp([0, L])$ as follows

$$V_k = \{F \in C^0_\sharp([0,L]) \mid F_{|[x_i,x_{i+1}]} \in \mathbb{P}_k\},$$
$$W_k = \{C \in L^2_\sharp([0,L]) \mid C_{|[x_i,x_{i+1}]} \in \mathbb{P}_{k-1}\},$$

where we denote by $\mathbb{P}_k$ the space of polynomials of degree less or equal to k. The functions of V_k are continuous and piecewise polynomials of degree k and the functions of W_k piecewise polynomials of degree $k-1$ with no continuity requirements at the grid points. The dimension of $\mathbb{P}_k$ the space of polynomials of one variable of degree less or equal to k is $k+1$. It follows that, due to the continuity requirement at the grid points, the dimension of V_k is Nk for N cells, and the dimension of W_0 is also Nk. Notice that the derivatives of functions of V_k are in W_k.

The discrete variational formulation in the spaces V_k and W_k then reads
Find $(E_h, B_h) \in V_k \times W_k$ *such that*

$$\frac{d}{dt}\int_0^L E_h(x,t)F(x)\,dx - c^2\int_0^L B_h(x,t)\frac{\partial F}{\partial x}(x)\,dx = -\frac{1}{\varepsilon_0}\int_0^L J(x,t)F(x)\,dx \quad \forall F \in V_k, \tag{3.20}$$

$$\frac{d}{dt}\int_0^L B_h(x,t)C(x)\,dx + \int_0^L \frac{\partial E_h}{\partial x}(x,t)C(x)\,dx = 0 \quad \forall C \in W_k. \tag{3.21}$$

We shall now express this variational formulations in finite dimensional spaces in matrix form using appropriate basis functions of the spaces V_k and W_k. In the case of high order methods we need to keep in mind that the condition number of the elementary mass matrices should be kept as low as possible in order to avoid problems coming from round-off errors. We shall consider here for simplicity only Lagrange Finite Elements, where the degrees of freedom are point values at Lagrange interpolating points.

It is well known in particular that Lagrange Finite Elements with uniformly distributed interpolation points (degrees of freedom) lead to very ill conditioned matrices for moderately large values of k. A much better option is to use Lagrange polynomials at Gauss points. This is the best choice for W_k. For V_k we have the additional continuity condition at the grid points which forces us to put degrees of freedom at the grid points. Then the best choice is the Gauss-Lobatto points.

Remark 5 *Note that if one does not want to stick to Lagrange Finite Elements a natural choice of basis functions for W_k would be the orthonormal Legendre polynomials for which the mass matrix would be identity.*

Let us denote by $(\varphi_i)_{0\leq kN-1}$ the basis of V_k and $(\psi_j)_{0\leq j\leq kN-1}$ the basis of W_k.

Let us know compute the different integrals appearing in the variational formulation (3.20)-(3.21) using the basis functions. Expressing E_h and F using the basis (φ_i) and B_h and C in the basis (ψ_j). We get

$$E_h(x,t) = \sum_{j=0}^{kN-1} E_j(t)\varphi_j(x), \quad F(x) = \sum_{i=0}^{kN-1} F_i\varphi_i(x),$$

$$B_h(x,t) = \sum_{j=0}^{kN-1} B_j(t)\psi_j(x), \quad C(x) = \sum_{i=0}^{kN-1} F_i\psi_i(x).$$

Note that as both bases are Lagrange bases the coefficients E_i, F_i, B_i, C_i are simply the values of the corresponding functions E_h, F, B_h and C at the corresponding points.

Plugging these expression into (3.20)-(3.21) we obtain

$$\sum_{i=0}^{kN-1}\sum_{j=0}^{kN-1}\left[\frac{dE_j(t)}{dt}F_i\int_0^L \varphi_i(x)\varphi_j(x)\,dx - B_j(t)F_i\int_0^L \varphi_i'(x)\psi_j(x)\,dx\right] \tag{3.22}$$

$$= \sum_{i=0}^{N-1} F_i\int_0^L J(x)\varphi_i(x)\,dx$$

$$\sum_{i=0}^{kN-1}\sum_{j=0}^{kN-1}\left[\frac{dB_j(t)}{dt}C_i\int_0^L \psi_i(x)\psi_j(x)\,dx + E_j(t)C_i\int_0^L \varphi_j'(x)\psi_i(x)\,dx\right] = 0. \tag{3.23}$$

Denote by

$$\mathbb{E}(t) = \begin{pmatrix} E_0(t) \\ \vdots \\ E_{N-1}(t)\end{pmatrix}, \quad \mathbb{B}(t) = \begin{pmatrix} B_0(t) \\ \vdots \\ B_{N-1}(t)\end{pmatrix}, \quad \mathbb{F} = \begin{pmatrix} F_0 \\ \vdots \\ F_{N-1}\end{pmatrix}, \quad \mathbb{C} = \begin{pmatrix} C_0 \\ \vdots \\ C_{N-1}\end{pmatrix},$$

and

$$\mathbb{J}(t) = \begin{pmatrix} \int_0^L J(t,x)\varphi_0(x)\,dx \\ \vdots \\ \int_0^L J(t,x)\varphi_{N-1}(x)\,dx\end{pmatrix}.$$

Let us now introduce the mass matrices

$$M_E = ((\int_0^L \varphi_i(x)\varphi_j(x)))_{0\leq i\leq Nk-1,0\leq j\leq Nk-1}, \quad M_B = ((\int_0^L \psi_i(x)\psi_j(x)))_{0\leq i\leq Nk-1,0\leq j\leq Nk-1},$$

and the derivative matrix

$$K = ((\int_0^L \varphi_j'(x)\psi_i(x)))_{0\leq i\leq Nk-1,0\leq j\leq Nk-1}.$$

The variational formulations (3.22)-(3.23) then become

$$\mathbb{F}^T M_E \frac{d\mathbb{E}(t)}{dt} - c^2 \mathbb{F}^T K^T \mathbb{B} = -\frac{1}{\varepsilon_0} \mathbb{F}^T \mathbb{J} \quad \forall \mathbb{F} \in \mathbb{R}^{N-1},$$

$$\mathbb{C}^T M_B \frac{d\mathbb{B}(t)}{dt} + \mathbb{C}^T K \mathbb{E} = 0 \quad \forall \mathbb{C} \in \mathbb{R}^{N-1}.$$

As M_E and M_B are non singular matrices this can be written equivalently

$$\frac{d\mathbb{E}(t)}{dt} - M_E^{-1} K^T \mathbb{B} = 0, \tag{3.24}$$

$$\frac{d\mathbb{B}(t)}{dt} + M_B^{-1} K \mathbb{E} = 0. \tag{3.25}$$

As usual for Finite Elements the matrices M_B, M_E and K are computed from the corresponding elementary matrices that are obtained by change of variables onto the reference element $[-1, 1]$ for each cell. So

$$\int_0^L \varphi_i(x)\varphi_j(x)\, dx = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varphi_i(x)\varphi_j(x)\, dx,$$

and doing the change of variable $x = \frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2}$, we get

$$\int_{x_n}^{x_{n+1}} \varphi_i(x)\varphi_j(x)\, dx = \frac{x_{n+1}-x_n}{2} \int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})\, d\hat{x},$$

where $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2})$. The local indices α on the reference element go from 0 to k and the global numbers of the basis functions not vanishing on element n are $j = kn + \alpha$. The $\hat{\varphi}_\alpha$ are the Lagrange polynomials at the Gauss-Lobatto points in the l'interval $[-1, 1]$.

The mass matrix in V_k can be approximated with no loss of order of the finite element approximation using the Gauss-Lobatto quadrature rule. Then because the products $\hat{\varphi}_\alpha(\hat{x})\hat{\varphi}_\beta(\hat{x})$ vanish for $\alpha \neq \beta$ at the Gauss-Lobatto points by definition of the $\hat{\varphi}_\alpha$ which are the Lagrange basis functions at these points, the elementary matrix $\hat{M}_E$ is diagonal and we have

$$\int_{-1}^{1} \hat{\varphi}_\alpha(\hat{x})^2\, d\hat{x} \approx \sum_{\beta=0}^{k} w_\beta^{GL} \varphi_\alpha(\hat{x}_\beta)^2 = w_\alpha^{GL}$$

using the quadrature rule, where w_α^{GL} is the Gauss-Lobatto weight at Gauss-Lobatto point $(\hat{x}_\alpha) \in [-1, 1]$. So that finally $\hat{M}_E = diag(w_0^{GL}, \dots w_k^{GL})$ is the matrix with $k+1$ lines and columns with the Gauss-Lobatto weights on the diagonal.

In the same spirit we use Gauss quadrature to compute the mass matrix in W_k. In this case, the quadrature is exact and we get in the same way an elementary matrix which is diagonal of order k where the diagonal terms are the Gauss weights: $\hat{M}_B = diag(w_0^G, \dots w_{k-1}^{GL})$.

Let us now compute the elements of K. As previously we go back to the interval $[-1, 1]$ with the change of variables $x = \frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2}$ and we define $\hat{\varphi}_\alpha(\hat{x}) = \varphi_i(\frac{x_{n+1}-x_n}{2}\hat{x} + \frac{x_{n+1}+x_n}{2})$ and the same for ψ_i. Note however that a global basis function φ_i associated to a grid point has a support which overlaps two cells and is associated to

two local basis functions whereas the global basis functions ψ_i have all only a support on one cell and are only associated to one local basis functions. This is the reason why there are $k+1$ local basis functions $\hat{\varphi}_\alpha$ and only k local basis functions $\hat{\psi}_\alpha$ even though there are in our case the same number of global basis functions for V_k and W_k.

We then get $\hat{\varphi}'_\alpha(\hat{x}) = \frac{x_{n+1}-x_n}{2}\varphi'_i(\frac{x_{n+1}-x_n}{2}(\hat{x}+1)+x_n)$. It follows that

$$\int_{x_n}^{x_{n+1}} \varphi'_j(x)\psi_i(x)\,dx = \int_{-1}^{1} \frac{2}{x_{n+1}-x_n}\hat{\varphi}'_\beta(\hat{x})\hat{\psi}_\alpha(\hat{x})\frac{x_{n+1}-x_n}{2}\,d\hat{x} = \int_{-1}^{1} \hat{\varphi}'_\beta(\hat{x})\hat{\psi}_\alpha(\hat{x})\,d\hat{x}.$$

Both $\hat{\varphi}'_\beta(\hat{x})$ and $\hat{\psi}_\alpha(\hat{x})$ are of degree $k-1$ so that the Gauss-Lobatto quadrature rule with $k+1$ points is exact. Using this rule

$$\int_{-1}^{1} \hat{\varphi}'_\beta(\hat{x})\hat{\psi}_\alpha(\hat{x})\,d\hat{x} = \sum_{m=0}^{k} w_m^{GL}\hat{\varphi}'_\beta(\hat{x}_m^{GL})\hat{\psi}_\alpha(\hat{x}_m^{GL}).$$

In this case there is no vanishing term, but this expression can be used to compute the elementary matrix $\hat{K}$ along with the formula for the Lagrange polynomial at the Gauss points

$$\hat{\psi}_\alpha(\hat{x}_m^{GL}) = \frac{\pi_{\beta\neq\alpha}(\hat{x}_m^{GL}-\hat{x}_\beta^G)}{\pi_{\beta\neq\alpha}(\hat{x}_\alpha^G-\hat{x}_\beta^G)}.$$

On the other hand evaluating the derivatives of the Lagrange polynomial at the Gauss-Lobatto points at these Gauss-Lobatto points can be done using the formula

$$\hat{\varphi}'_\alpha(\hat{x}_\beta^{GL}) = \frac{p_\beta/p_\alpha}{\hat{x}_\beta^{GL}-\hat{x}_\alpha^{GL}} \text{ for } \beta\neq\alpha \text{ and } l'_\alpha(\hat{x}_\alpha) = -\sum_{\beta\neq\alpha} l'_\beta(\hat{x}_\alpha).$$

Note that the support of a function ψ_j is restricted to only one cell $[x_n, x_{n+1}]$, of the mesh. Therefore matrix K is consists in blocks of size $k\times(k+1)$ with only one block by group of k lines and with a common column corresponding to a grid point for two successive blocks.

3.1.5 B-spline Finite Elements

Let us now construct a different kind of Finite Element discretization using B-Splines as basis functions.

In order to define a family of n B-splines of degree k, we need $(x_i)_{0\leqslant i\leqslant n+k}$ a non-decreasing sequence of points on the real line called *knots* in the spline terminology. There can be several knots at the same position. In the case when there are m knots at the same point, we say that the knot has multiplicity m.

Definition 4 (B-Spline) *Let $(x_i)_{0\leqslant i\leqslant n+k}$ be a non-decreasing sequence of knots. Then the j-th B-Spline ($0\leq j\leq n-1$) denoted by N_j^k of degree k is defined by the recurrence relation:*

$$N_j^k(x) = w_j^k(x)N_j^{k-1}(x) + (1-w_{j+1}^k(x))N_{j+1}^{k-1}(x)$$

where,

$$w_j^k(x) = \frac{x-x_j}{x_{j+k}-x_j} \qquad\qquad N_j^0(x) = \chi_{[x_j,x_{j+1}[}(x)$$

We note some important properties of a B-splines basis:

- B-splines are piecewise polynomial of degree k,
- B-splines are non negative
- Compact support; the support of N_j^k is contained in $[t_j, .., t_{j+k+1}]$
- Partition of unity: $\sum_{i=0}^{n-1} N_i^k(x) = 1, \forall x \in \mathbb{R}$
- Local linear independence
- If a knot x_i has a multiplicity m then the B-spline is $\mathcal{C}^{(k-m)}$ at x_i.

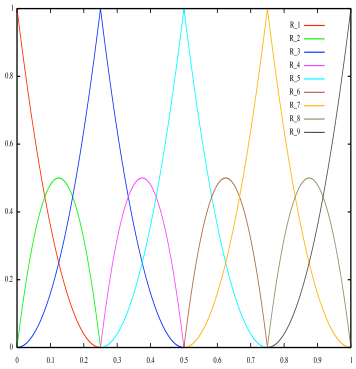

Figure 3.1: All B-splines functions associated to a knot sequence defined by $n = 9$, $k = 2$, $T = \{000, \frac{1}{4}\frac{1}{4}, \frac{1}{2}\frac{1}{2}, \frac{3}{4}\frac{3}{4}, 111\}$

A key point for constructing discrete Finite Element spaces for the Maxwell equation comes from the recursion formula for the derivatives:

$$N_i^{k\prime}(x) = k\left(\frac{N_i^{k-1}(x)}{x_{i+k} - x_i} - \frac{N_{i+1}^{k-1}(x)}{x_{i+k+1} - x_{i+1}}\right). \tag{3.26}$$

It will be convenient to introduce the notation $D_i^k(x) = k\frac{N_i^{k-1}(x)}{x_{i+k}-x_i}$. Then the recursion formula for derivative simply becomes

$$N_i^{k\prime}(x) = D_i^k(x) - D_{i+1}^k(x). \tag{3.27}$$

Remark 6 *In the case where all knots, except the boundary knots are of multiplicity 1, the set $(N_i^k)_{0\leq i\leq n-1}$ of B-splines of degree k forms a basis of the spline space defined by*

$$\mathcal{S}^k = \{v \in C^{k-1}([x_0, x_n]) \; | v_{|[x_i,x_{i+1}]} \in \mathbb{P}_k([x_i, x_{i+1}]).$$

The boundary knots are chosen to have multiplicity $k+1$ so that the spline becomes interpolatory on the boundary in order to simplify the application of Dirichlet boundary conditions.

Then due to the definitions if follows immediately that $(D_i^k)_{1\leq i\leq n-1}$ is a basis of $\mathcal{S}^{k-1}$. Note that if the first knot has multiplicity $k+1$, D_0^k will have a support restricted to one point and be identically 0.

Remark 7 *Splines can be easily defined in the case of periodic boundary conditions by taking a periodic knot sequence.*

Assuming only knots of multiplicity 1 and denoting by $\mathcal{S}_\sharp^k$ the set of periodic splines associated to a periodic knot sequences, we can take $V_k = \mathcal{S}_\sharp^k$ whose basis functions are the N_i^k and $W_k = \mathcal{S}_\sharp^{k-1}$ with basis functions the D_i^k. This defines the Finite Element spaces that we can use with the discrete variational formulation of Maxwell's equations (3.20)-(3.21). We can then construct the mass and derivative matrices like for the Lagrange Finite Elements, just replacing the basis functions by their spline counterparts, *i.e.* φ_i by N_i^k and ψ_i by D_i^k. The matrices can be computed with no quadrature error using adequate Gauss or Gauss-Lobato formulas. We then get a linear system which has the form (3.24)-(3.25). Note that in this case the mass matrices will not become diagonal thanks to the quadrature formula. However, as one of the two variational formulations, the Faraday equation (3.21) in our case (but it could have been the other one), has not been integrated by part it can be written in strong form in the discretization space thanks to the choice of the spaces. Indeed, we have that the derivatives of the functions of V_k are in W_k. Hence $\frac{\partial E_h}{\partial x} \in W_k$ and (3.21) implies that $\frac{\partial B_h}{\partial t} + \frac{\partial E_h}{\partial x}$ is a function of W_k which is orthogonal to all functions of W_k and thus vanishes. The second variational formulation is thus equivalent to the strong form

$$\frac{\partial B_h}{\partial t} + \frac{\partial E_h}{\partial x} = 0.$$

Let us now write the expressions of B_h and E_h in their respective basis: $B_h = \sum_i \mathbb{B}_i(t) D_i^k(x)$ and $E_h = \sum_i \mathbb{E}_i(t) N_i^k(x)$. Note that unlike for Lagrange finite elements, the coefficients $\mathbb{B}_i(t)$ and $\mathbb{E}_i(t)$ on the bases are not values of the discrete functions at some given points. Then taking the derivative with respect to x of E_h and using the B-spline derivative formula (3.27) we find

$$\frac{\partial E_h}{\partial x} = \sum_i \mathbb{E}_i(t)(N_i^k)'(x) = \sum_i \mathbb{E}_i(t)(D_i^k(x) - D_{i+1}^k(x)) = \sum_i (\mathbb{E}_i(t) - \mathbb{E}_{i-1}) D_i^k(x).$$

Hence identifying the coefficients of the basis, the discrete Faraday equation reduces in this case to

$$\frac{d\mathbb{B}_i}{dt} = \mathbb{E}_i(t) - \mathbb{E}_{i-1}(t).$$

On the other hand, as in the case of the mixed Lagrange Finite Elements the discrete matrix form of Ampère's law writes

$$M_V \frac{d\mathbb{E}}{dt} = K^T \mathbb{B},$$

where $M_V = ((\int N_i^k(x) N_j^k(x)\, dx))_{0 \le i,j \le n-1}$ and $K = ((\int (N_i^k)'(x) D_j^k(x)\, dx))_{0 \le i,j \le n-1}$.

3.1.6 Variationnal formulations for the 2D Maxwell equations

In order to introduce the Finite Element formulation, we need the variational form of Maxwell's equations (3.10)-(3.12). The unknowns of Maxwell's equations live in related function spaces.

Indeed in two dimensions we have the following diagrams, which are called *exact sequences*:

$$H^1(\Omega) \stackrel{grad}{\longrightarrow} H(\mathrm{curl}\,,\Omega) \stackrel{\mathrm{curl}}{\longrightarrow} L^2(\Omega)$$

and

$$H(\mathbf{curl}\,,\Omega) \stackrel{\mathbf{curl}}{\longrightarrow} H(\mathrm{div}\,,\Omega) \stackrel{div}{\longrightarrow} L^2(\Omega).$$

Note that $\mathbf{curl}\,\varphi \in L^2(\Omega)^2$ is equivalent to $\nabla\varphi \in L^2(\Omega)^2$ and hence the spaces $H(\mathbf{curl}\,,\Omega)$ and $H^1(\Omega)$ are identical.

The exact sequences mean that the image of the space on the left hand side of the arrow by the operator on top of the arrow is included in the space on the right hand side of the arrow and moreover that it is equal to the kernel of the following operator.

In order to conserve at the discrete level the properties of the continuous equations, it is necessary to look for the unknowns in the spaces associated to one of the exact sequence and introduce discrete spaces satisfying the same exact sequence property. This will be convenient in practice as one of Ampère's law or Faraday's law can be kept in strong form removing the need for inverting a mass matrix. Moreover this setting enables to prove stability and convergence in a very general manner [1, 2].

Thus, we have two choices for the variational form, either to use a weak form of Ampère's law and keep Faraday's law as it is or the opposite. The first option then consists in looking for $\mathbf{E} \in H(\mathrm{curl}\,,\Omega)$ and $B_z \in L^2(\Omega)$ using the last arrow of the first diagram and the second option in looking for $\mathbf{E} \in H(div,\Omega)$ and $B_z \in H^1(\Omega) = H(\mathbf{curl}\,,\Omega)$ using the first arrow of the second diagram. Let us consider here the first option.

We shall need the following 2D Green formulae to derive the variational formulations:

$$\int_\Omega \mathbf{F}\cdot\mathbf{curl}\,C - \int_\Omega \mathrm{curl}\,\mathbf{F}\,C = \int_{\partial\Omega} (\mathbf{F}\cdot\boldsymbol{\tau})\,C \quad \forall \mathbf{F} \in H(\mathrm{curl}\,,\Omega),\ C \in H^1(\Omega), \tag{3.28}$$

$$\int_\Omega \mathrm{div}\,\mathbf{F}\,q + \int_\Omega \mathbf{F}\cdot\nabla q = \int_{\partial\Omega} (\mathbf{F}\cdot\mathbf{n})\,q \quad \forall \mathbf{F} \in H(\mathrm{div}\,,\Omega),\ q \in H^1(\Omega). \tag{3.29}$$

We shall look for $\mathbf{E} \in H(\mathrm{curl}\,,\Omega)$ and $B_z \in L^2(\Omega)$. In order to get the weak form, we take the scalar product of (3.10) with a test function $\mathbf{F} \in H(curl,\Omega)$ and integrate on Ω

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} - c^2\int_\Omega \mathbf{curl}\,B_z\cdot\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F}.$$

Using the Green formula (3.28) this becomes, assuming the boundary term vanishes, which is the case for perfect conductor or periodic boundary conditions

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\mathbf{F} + c^2\int_\Omega B_z\,\mathrm{curl}\,\mathbf{F} = -\frac{1}{\varepsilon_0}\int_\Omega \mathbf{J}\cdot\mathbf{F} \quad \forall \mathbf{F} \in H(\mathrm{curl}\,,\Omega). \tag{3.30}$$

For Faraday's law, we multiply by $C \in L^2(\Omega)$ and integrate on Ω which yields

$$\frac{d}{dt}\int_\Omega B_z\,C + \int_\Omega \mathrm{curl}\,\mathbf{E}\,C = 0 \quad \forall C \in L^2(\Omega). \tag{3.31}$$

For Gauss's law, we multiply by $q \in H^1(\Omega)$ and integrate on Ω which yields

$$\int_\Omega \mathrm{div}\,\mathbf{E}\,q = \frac{1}{\varepsilon_0}\int_\Omega \rho\,q.$$

And using (3.29) and assuming that $\mathbf{E} \cdot \mathbf{n} = 0$ on the boundary or periodic boundary conditions, we obtain that

$$-\int_\Omega \mathbf{E} \cdot \nabla q = \frac{1}{\varepsilon_0} \int_\Omega \rho\, q \quad \forall q \in H^1(\Omega). \tag{3.32}$$

Finally, let us also write the weak form of the continuity equation (3.13), that we multiply by $q \in H^1(\Omega)$ and integrate using (3.29) and the fact that $\mathbf{J} \cdot \mathbf{n} = 0$ on the boundary or periodic boundary conditions, like in the previous case. We then get

$$\frac{d}{dt} \int_\Omega \rho\, q = \int_\Omega \mathbf{J} \cdot \nabla q \quad \forall q \in H^1(\Omega). \tag{3.33}$$

Remark 8 *Let us note that if the weak form of Gauss' law (3.32) is verified at $t = 0$ and if the weak continuity equation (3.33) is verified, the weak form of Gauss' law is verified for all times. Indeed, for any $q \in H^1(\Omega)$, $\operatorname{curl} \nabla q = 0$ and hence $\nabla q \in H(\operatorname{curl}, \Omega)$, then we can use ∇q as a test function in* (3.30)*, which becomes*

$$\frac{d}{dt} \int_\Omega \mathbf{E} \cdot \nabla q = -\int_\Omega \mathbf{J} \cdot \nabla q \quad \forall q \in H^1(\Omega).$$

Then, using (3.33),

$$\frac{d}{dt}\Big(\int_\Omega \mathbf{E} \cdot \nabla q + \int_\Omega \rho\, q\Big) \quad \forall q \in H^1(\Omega),$$

which gives the desired result.

Let us now introduce the second variational formulation which involves $\mathbf{E} \in H(div, \Omega)$ and $B_z \in H^1(\Omega) (= H(\mathbf{curl}, \Omega))$. In order to get the variationl form, we take the scalar product of (3.10) with a test function $\mathbf{F} \in H(div, \Omega)$ and integrate on Ω

$$\frac{d}{dt} \int_\Omega \mathbf{E} \cdot \mathbf{F} - c^2 \int_\Omega \mathbf{curl}\, B_z \cdot \mathbf{F} = -\frac{1}{\varepsilon_0} \int_\Omega \mathbf{J} \cdot \mathbf{F}, \quad \forall \mathbf{F} \in H(div, \Omega). \tag{3.34}$$

For Faraday's law, we multiply by $C \in L^2(\Omega)$ and integrate on Ω which yields

$$\frac{d}{dt} \int_\Omega B_z\, C + \int_\Omega \operatorname{curl} \mathbf{E}\, C = 0,$$

Using the Green formula (3.28) this becomes

$$\frac{d}{dt} \int_\Omega B_z\, C + \int_\Omega \mathbf{E} \cdot \mathbf{curl}\, C = 0 \quad \forall C \in H^1(\Omega). \tag{3.35}$$

3.1.7 Discretization using conforming finite elements

In order to keep the specific features of Maxwell's equations at the discrete level which are useful in different contexts, we shall consider finite dimensional subspaces endowed with the same exact sequence structure as in the continuous level [2, 1, 10]. Let us first derive the linear system that comes out of this discretization. Let $\{\boldsymbol{\psi}_i\}_{i=1\ldots N}$ be a basis of $W \subset H(\operatorname{curl}, \Omega)$ and $\{\varphi_k\}_{k=1\ldots M}$ a basis of $V \subset L^2(\Omega)$. Then denoting by σ_i^W the degrees of freedom associated to $\{\boldsymbol{\psi}_i\}_{i=1\ldots N}$ and by σ_k^V the degrees of freedom associated

to $\{\varphi_k\}_{k=1\dots M}$. In the case of Lagrange Finite Elements, this degrees of freedom are just point values. We can write elements of W and V respectively

$$\mathbf{E}_h = \sum_{i=1}^{N} \sigma_i^W(\mathbf{E}_h)\boldsymbol{\psi}_i, \qquad B_h = \sum_{k=1}^{M} \sigma_k^V(B_h)\varphi_k.$$

Replacing the continuous spaces by the discrete spaces in the variational formulations (3.30) and (3.31) we get the following discrete problem:
Find $(\mathbf{E}_h, B_h) \in W \times V$ such that

$$\frac{d}{dt}\int_\Omega \mathbf{E}\cdot\boldsymbol{\psi_i}\, d\mathbf{x} - \int_\Omega B\,(\operatorname{curl}\boldsymbol{\psi}_i)\, d\mathbf{x} = -\int_\Omega \mathbf{J}\cdot\boldsymbol{\psi}_i\, d\mathbf{x}, \qquad \forall i = 1\dots N, \tag{3.36}$$

$$\frac{d}{dt}\int_\Omega B\,\varphi_k\, d\mathbf{x} + \int_\Omega (\operatorname{curl}\mathbf{E})\,\varphi_k\, d\mathbf{x} = 0, \qquad \forall k = 1\dots M, \tag{3.37}$$

which becomes when $\mathbf{E}$ and B are decomposed on the respective bases of W and V

$$M_W\frac{dE}{dt} - KB = J \tag{3.38}$$

$$M_V\frac{dB}{dt} + K^T E = 0 \tag{3.39}$$

where $E = (\sigma_i^W(\mathbf{E}_h))_{1\le i\le N}$ (resp. $B = (\sigma_k^V(B_h))_{1\le k\le M}$) denote vectors of degrees of freedom for the discrete electric and magnetic fields, with

$$(M_W)_{1\le i,j\le N} = \int_\Omega \boldsymbol{\psi}_j\cdot\boldsymbol{\psi}_i\, d\mathbf{x},\ (M_V)_{1\le i,j\le M} = \int_\Omega \varphi_i\varphi_j\, d\mathbf{x},$$

$$(K)_{1\le i\le N,\,1\le j\le M} = \int_\Omega \varphi_j(\operatorname{curl}\boldsymbol{\psi}_i)\, d\mathbf{x}.$$

Remark 9 *Notice that the structure of this linear system is exactly the same as in the 1D case. Hence the same time schemes with the same properties can be used.*

As we already noticed in the 1D case, the exact sequence structure of our Finite Element spaces enables us to express the discrete variational formulation where no Green formula (or integration by parts) was used (3.37) in our case by a strong form. Indeed the exact sequence structure of our discrete spaces implies in particular that if $\mathbf{E}_h \in W$, we have $\operatorname{curl}\mathbf{E}_h \in V$. Hence we can express $\operatorname{curl}\mathbf{E}_h$ on the basis $\{\varphi_k\}_{k=1,\dots,M}$ of V which yields

$$\operatorname{curl}\mathbf{E}_h = \sum_{l=1}^{M}\sigma_l^V(\operatorname{curl}\mathbf{E}_h)\varphi_l = \sum_{l=1}^{M}\sigma_l^V\Big(\sum_{j=1}^{N}\sigma_j^W(\mathbf{E}_h)(\operatorname{curl}\boldsymbol{\psi}_j)\Big)\varphi_l = \sum_{l=1}^{M}\sum_{j=1}^{N}\sigma_j^W(\mathbf{E}_h)\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)\varphi_l.$$

In particular we get that $\sigma_l^V(\operatorname{curl}\mathbf{E}_h) = \sum_{j=1}^{N}\sigma_j^W(\mathbf{E}_h)\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)$, and injecting this expression in the discrete Faraday law

$$\begin{aligned}\int_\Omega (\operatorname{curl}\mathbf{E})\varphi_k\, d\mathbf{x} &= \int_\Omega \sum_{l=1}^{M}\sum_{j=1}^{N}\sigma_j^W(\mathbf{E})\sigma_l^V(\operatorname{curl}\boldsymbol{\psi_j})\varphi_l\varphi_k\, d\mathbf{x}\\ &= \sum_{l=1}^{M}\sum_{j=1}^{N}\int_\Omega \varphi_l\varphi_k\, d\mathbf{x}\,\sigma_l^V(\operatorname{curl}\boldsymbol{\psi}_j)\,\sigma_j^W(\mathbf{E}),\end{aligned}$$

which is the k^{th} line (for k from 1 to M) of the vector $M_V RE$ where R is the matrix defined by

$$(R)_{1\leq i\leq M,\,1\leq j\leq N} = \sigma_i^V(\operatorname{curl}\boldsymbol{\psi}_j),$$

so that, as M_V is non singular, the system (3.38)-(3.39) is algebraically equivalent to the system

$$M_W\frac{dE}{dt} - KB = J, \tag{3.40}$$

$$\frac{dB}{dt} + RE = 0, \tag{3.41}$$

This new formulation yields an explicit expression of B which can then be computed without solving a linear system.

Now having expressed the general structure that we want our Finite Element spaces to have and its consequences, there is still a wide variety of choices of Finite Element spaces that verify these exact sequence property. Let us define some classical spaces on quads and triangles.

On rectangular meshes the cells of which are denoted by K_i, $1 \leq i \leq r$, the three actual subspaces of the exact sequence $X \subset H^1(\Omega)$, $W \subset H(\operatorname{curl},\Omega)$ and $V \subset L^2(\Omega)$ can be defined as follows

$$X = \{\chi \in H^1(\Omega) \;\mid\; \chi_{|K_i} \in \mathbb{Q}_k(K_i)\},$$

$$W = \{\boldsymbol{\psi} \in H(\operatorname{curl},\Omega) \;\mid\; \boldsymbol{\psi}_{|K_i} \in \begin{pmatrix} \mathbb{Q}_{k-1,k}(K_i) \\ \mathbb{Q}_{k,k-1}(K_i) \end{pmatrix},\ \forall i = 1,\ldots,r\},$$

$$V = \{\varphi \in L^2(\Omega) \;\mid\; \varphi_{|K_i} \in \mathbb{Q}_{k-1}(K_i)\ ,\ \forall i = 1,\ldots,r\}.$$

where $\mathbb{Q}_{m,n} = \left\{x^i y^j\ ,\ 0 \leq i \leq m, 0 \leq j \leq n\right\}$, with the particular case $\mathbb{Q}_{m,m}$ is simply denoted by $\mathbb{Q}_m$ in the classical way.

All these Finite Element spaces are piecewise polynomials for scalar fields in the case of X and V and for each component of a field for W. In the case of V which is just in $L^2(\Omega)$ there is no additional continuity requirement at the cell interface. In the case of X the inclusion in $H^1(\Omega)$ imposes continuity at the cell interface and in the case of W the inclusion in $H(\operatorname{curl},\Omega)$ imposes continuity of the tangential component of the field at the cell interface.

The space $\mathbb{Q}_k$ is the standard continuous Lagrange Finite Element space on quads. The space W is known as the first family of edge elements $H(\operatorname{curl})$-conforming of Nédélec [13] and the space V is the space of discontinuous functions which restrict to a polynomial of degree $k-1$ with respect to each variable on each cell. This is the kind of approximation used in Discontinuous Galerkin methods.

After having defined the spaces there are still many choices for the degrees of freedom which define the actual basis functions. In the interpretation of Maxwell's equations as differential forms it is natural to take the degrees of freedom for X which corresponds to 0-forms as point values, the degrees of freedom for W which corresponds to 1-forms as edge integrals, and the degrees of freedom for V which corresponds to 2-forms as cell integrals. But such a choice is not mandatory. The Cohen Monk Finite Elements for W are based on Lagrange degrees of freedom (point values) for Maxwell at Gauss or Gauss-Lobatto points. This has the advantage of leading to a diagonal mass matrix M_W if the mesh is cartesian thanks to the Gauss-Lobatto quadrature formula [4, 5].

Let us now introduce in detail the degrees of freedom that are obtained by an interpretation in terms of differential forms, which have very convenient properties. For this, two types of 1D basis functions are needed in a tensor product construction on quads, first the nodal basis functions which typically are the standard Lagrange basis functions and then the edge basis functions, which are constructed from the nodal basis functions and whose degrees of freedom are edge integrals. Let us consider the 1D reference element $[-1, 1]$ on which we define the Gauss-Lobatto points $(\hat{x}_i)_{0\leq i\leq k}$. Note that uniform interpolation points could also be used for the construction, but they are not stable for higher degrees. We denote by $l_{k,i}$, $0 \leq i \leq k$ the Lagrange basis functions associated to these points. This will define the local basis, on each element, of the discrete space X. This is a standard Lagrange Finite Element for which the degrees of freedom are the point values at the interpolation points and we have $l_{k,i}(\hat{x}_j) = \delta_{i,j}$.

Our aim is now to use this Lagrange basis to construct a basis e_i $0 \leq i \leq k-1$ for W the next space in the sequence. This should be such that the derivatives of linear combinations of the Lagrange basis functions are exactly represented. It will be natural to define the degrees of freedom of this space to be integrals between two successive interpolation points, so that the degrees of freedom of a derivative $u = \frac{d\phi}{dx}$ can be expressed directly with respect to the degrees of freedom of ϕ in X. Indeed

$$\int_{\hat{x}_\nu}^{\hat{x}_{\nu+1}} \frac{d\phi}{d\hat{x}}(\hat{x})\,\mathrm{d}\hat{x} = \phi(\hat{x}_{\nu+1}) - \phi(\hat{x}_\nu).$$

Next we find that the basis associated to these degrees of freedom, *i.e.* verifying

$$\int_{\hat{x}_{j-1}}^{\hat{x}_j} e_i(\hat{x})\,d\hat{x} = \delta_{i,j}, \quad 1 \leq j \leq k.$$

This can be expressed (see [8]) by

$$e_i(\hat{x}) = -\sum_{\nu=0}^{i-1} \frac{dl_{k,\nu}}{d\hat{x}}(\hat{x}), \quad 1 \leq j \leq k. \tag{3.42}$$

Indeed, we have for $1 \leq i, j \leq k$

$$\begin{aligned}\int_{\hat{x}_{j-1}}^{\hat{x}_j} e_i(\hat{x})\,d\hat{x} = -\sum_{\nu=0}^{i-1}\int_{\hat{x}_{j-1}}^{\hat{x}_j} \frac{dl_{k,\nu}}{d\hat{x}}(\hat{x})\,d\hat{x} &= -\sum_{\nu=0}^{i-1}(l_{k,\nu}(\hat{x}_j) - l_{k,\nu}(\hat{x}_{j-1})) \\ &= -\sum_{\nu=0}^{i-1}(\delta_{\nu,j} - \delta_{\nu+1,j}) = -(\delta_{0,j} - \delta i, j) = \delta_{i,j}\end{aligned}$$

as $\delta_{0,j} = 0$ for all $1 \leq j \leq k$.

Now having the local 1D basis functions $(l_{k,i})_{0\leq i\leq k}$ and $(e_i)_{1\leq i\leq k}$, the local 2D basis functions are defined using products of this basis functions.

The local basis functions defining X are the classical $\mathbb{Q}_k$ basis functions $l_k(x, y) = l_{k,i}(x)l_{k,j}(y)$ and the degrees of freedom the values at the points $\hat{x}_i\hat{y}_j$ $0 \leq i, j \leq k$, where the $\hat{x}_i$ as well as the $\hat{y}_j$ are the $k+1$ Gauss-Lobatto points.

Let us denote by $\mathbf{P}$ the set of local basis functions on which W is build

$$\mathbf{P} = \left\{ \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \text{ with } p_1 \in \mathbb{Q}_{k-1,k}, p_2 \in \mathbb{Q}_{k,k-1} \right\}.$$

Hence the basis functions are of two forms

$$\mathbf{e}^1_{i,j} = \begin{pmatrix} e_i(x) l_{k,j}(y) \\ 0 \end{pmatrix},\ 1 \le i \le k, 0 \le j \le k \quad \mathbf{e}^2_{i,j} = \begin{pmatrix} 0 \\ l_{k,i}(x) e_j(y) \end{pmatrix},\ 0 \le i \le k, 1 \le j \le k.$$

The degrees of freedom associated to the first set of basis functions are of the form

$$\mathbf{p} \mapsto \int_{\hat{x}_{i-1}}^{\hat{x}_i} p_1(x, y_j)\, dx \quad 1 \le i \le k, 0 \le j \le k,$$

and the degrees of freedom associated to the second set of basis functions are of the form

$$\mathbf{p} \mapsto \int_{\hat{y}_{j-1}}^{\hat{y}_j} p_2(x_i, y)\, dy \quad 0 \le i \le k, 1 \le j \le k.$$

Finally for V the local polynomial space is $\mathbf{Q}_{k-1}$ and the degrees of freedom are cell integrals. Hence the local basis functions can be expressed as $s(x, y) = e_i(x) e_j(y)$ $1 \le i, j \le k$ and the corresponding degrees of freedom are defined by

$$p \in \mathbf{Q}_{k-1} \mapsto \int_{\hat{x}_{i-1}}^{\hat{x}_i} \int_{\hat{y}_{j-1}}^{\hat{y}_j} p(x, y)\, dx\, dy.$$

These compatible Finite Element spaces enable to express the strong form of one of the Maxwell's equations in a very simple for directly relating the degrees of freedom using only the connectivity of the mesh and no geometric information like distances, lengths or areas. It is not modified by a compatible mapping.

On more general quad meshes the Finite Elements are defined via a bilinear transformation from a rectangular reference element associated to the spaces defined for one single element in the rectangular case. This is straightforward for the spaces of scalar fields X and V. But care must be taken for the space of vector fields W for which the vector valued basis functions need to be transformed in a covariant way to preserve the inclusion in $H(\mathrm{curl}, \Omega)$ of the discrete space W.

On triangular cells the discrete spaces are defined by

$$W = \{\boldsymbol{\psi} \in H(\mathrm{curl}, \Omega) \ \mid \ \boldsymbol{\psi}_{|T_i} \in \mathbb{P}^2_{k-1}(T_i) + \overline{\mathbb{P}}_{k-1}(T_i) \begin{pmatrix} y \\ -x \end{pmatrix},\ \forall i = 1, \ldots, r\},$$

$$V = \{\varphi \in L^2(\Omega) \ \mid \ \varphi_{|T_i} \in \mathbb{P}_{k-1}(T_i),\ \forall i = 1, \ldots, r\},$$

where $\overline{\mathbb{P}}_{k-1}$ denotes the set of polynomials of degree exactly $k-1$. The space V is $\mathbb{P}_{k-1}$ on each element and discontinuous across element boundaries (conforming in $L^2(\Omega)$), so is straightforward to construct. For the space W, we have again used the first family of edge elements of Nédélec [13], conforming in $H(\mathrm{curl}, \Omega)$, but we have changed the degrees of freedom.

3.2 The discontinuous Galerkin method

3.2.1 The Riemann problem for a 1D linear system

A general linear system of conservation laws in 1D can be written in the general form

$$\frac{\partial U}{\partial t} + A \frac{\partial U}{\partial x} = 0,$$

where $U(t,x)$ is a vector in $\mathbb{R}^n$ and A a given matrix with constant coefficients. We will focus in the following on the Discontinuous Galerkin method, which includes the first order Finite Volume method when polynomials of degree 0, i.e. constants are taken on each cell.

The main numerical issue when constructing a Discontinuous Galerkin scheme is to find a good numerical flux that is consistent (i.e. converges towards the exact flux when the cell size goes to 0) and stable. As we saw previously in the linear scalar case enhanced stability is given by upwinding. We now need to generalise the idea of upwind to the case of systems.

The construction of a numerical flux is a local procedure at the interface between two cells, where a different value is given on the left side and on the right side from the polynomial approximation (or reconstruction for Finite Volumes). In order to get information from the equation itself the idea is to solve it locally using an initial condition which is a step function. The Riemann problem is the corresponding initial value problem:

$$\begin{aligned} &\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0, \\ &U(0,x) = \left| \begin{array}{lll} U_L & \textit{if} & x < 0, \\ U_R & \textit{if} & x \geq 0, \end{array} \right. \end{aligned}$$

where U_L and U_R are two given constant vectors.

The system being hyperbolic implies that A has real eigenvalues and can be diagonalised. Hence $A = P\Lambda P^{-1}$, where Λ is the diagonal matrix containing the eigenvalues. Then introducing the so-called characteristic variables $V = P^{-1}U$, and multiplying the system by P^{-1} on the left we get

$$P^{-1}\frac{\partial U}{\partial t} + P^{-1}APP^{-1}\frac{\partial U}{\partial x} = \frac{\partial V}{\partial t} + \Lambda\frac{\partial V}{\partial x} = 0.$$

So in the variable V the system is diagonal and reduces to the set of linear advection equations

$$\frac{\partial v_i}{\partial t} + \lambda_i\frac{\partial v_i}{\partial x} = 0, \quad 1 \leq i \leq n$$

where the v_i are the components of V and the λ_i the eigenvalues of A. The exact solution of these equations is given by $v_i(t,x) = v_i(0, x - \lambda_i t)$, where the $v_i(0,x)$ are the components of the initial vector which take the constant values $V_L = P^{-1}U_L$ if $x < 0$ and $V_R = P^{-1}U_R$ if $x \geq 0$. In other terms

$$v_i(t,x) = \left| \begin{array}{lll} v_{i,L} & \textit{if} & x < \lambda_i t, \\ v_{i,R} & \textit{if} & x \geq \lambda_i t. \end{array} \right.$$

In practice we want to use the Riemann problem to determine the value of V (and U) at the cell interface, corresponding to $x = 0$, the discontinuity point at any strictly positive time. And we deduce from the previous solution that

$$v_i(t,0) = \left| \begin{array}{lll} v_{i,L} & \textit{if} & 0 < \lambda_i, \\ v_{i,R} & \textit{if} & 0 \geq \lambda_i. \end{array} \right.$$

In order to get a vector expression, we introduce the diagonal matrices Λ_+ where the negative eigenvalues are replaced by 0 and Λ_- where the positive eigenvalues are replaced by 0. Obviously $\Lambda = \Lambda_+ + \Lambda_-$. Then for $t > 0$ we have

$$\Lambda V(t,0) = \Lambda_+ V(t,0) + \Lambda_- V(t,0) = \Lambda_+ V_L + \Lambda_- V_R,$$

as for all positive eigenvalues the corresponding component of $V(t,0)$ is $v_{i,L}$ and for all negative eigenvalues the corresponding component of $V(t,0)$ is $v_{i,R}$. Note that as $V(t,0)$ is multiplied by Λ the components of $V(t,0)$ corresponding to 0 eigenvalues do not need to be considered as they are multiplied by 0 anyway. So the side where the strict inequality is used for the initial condition of the Riemann problem plays no role.

Denoting by $A_+ = P\Lambda_+P^{-1}$ and $A_- = P\Lambda_-P^{-1}$ the flux $AU(t,0)$ associated to the solution of the Riemann problem at the cell interface can also be expressed conveniently directly in terms of U

$$AU(t,0) = P\Lambda_+V(t,0) + P\Lambda_-V(t,0) = P\Lambda_+V_L + P\Lambda_-V_R = A_+U_L + A_-U_R.$$

This expression $AU(t,0) = A_+U_L + A_-U_R$ can be used to define the numerical flux at the cell interface, using the value U_L coming from the left-hand side of the interface and U_R coming from the right-hand side of the interface. For actual computations, the matrices A_+ and A_- need to be computed explicitely from the eigenvalues and eigenvectors of the matrix A. Notice that in the case of a scalar equation the matrix A is reduced to the scalar a which is then obviously the only eigenvalue of the 1×1 matrix and if $a > 0$ we have $A_+ = a$ and $A_- = 0$, so that the numerical flux becomes $au(t,0) = au_L$ and the same way if $a < 0$ $au(t,0) = au_R$, so that the numerical flux obtained from the solution of the Riemann problem reduces to the upwind flux.

Example. We consider the 1D Maxwell equations which can be written in dimensionless units:

$$\begin{aligned}\frac{\partial E}{\partial t} + \frac{\partial B}{\partial x} &= 0,\\ \frac{\partial B}{\partial t} + \frac{\partial E}{\partial x} &= 0.\end{aligned}$$

This can be written in the form of a linear system

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0, \quad \text{with } U = \begin{pmatrix} E \\ B \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The eigenvalues of A are the solutions of $\det(A - \lambda I) = 0$, *i.e.* $\lambda^2 = 1$. So the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 1$. They are real and distinct so that the system is strictly hyperbolic. Let V_i be a normalised eigenvector associated to the eigenvalue λ_i, $i = 1, 2$. We have $AV_1 = -V_1$ so that $V_1 = \frac{1}{\sqrt{2}}(1,-1)^T$ and $AV_2 = V_2$ so that $V_1 = \frac{1}{\sqrt{2}}(1,1)^T$. We define P the matrix whose columns are V_1 and V_2. P is obviously orthonormal, so that its inverse is its transpose. Then we have $PA = \Lambda P$. So that we can define:

$$A_+ = P\Lambda_+P^T = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

$$A_- = P\Lambda_-P^T = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Hence, the upwind flux is given by

$$AU(t,0) = A_+U_L + A_-U_R = \frac{1}{2}\begin{pmatrix} U_{L,1} + U_{L,2} + (-U_{R,1} + U_{R,2}) \\ U_{L,1} + U_{L,2} + (U_{R,1} - U_{R,2}) \end{pmatrix}.$$

Remark 10 *As for the scalar problem, the numerical flux can be taken as a linear combination of the centred flux and the upwind flux (solution the Riemann problem):*

$$G_j = \mu \frac{1}{2} A(U_L + U_R) + (1-\mu)(A_+ U_L + A_- U_R), \quad 0 \leq \mu \leq 1.$$

3.2.2 Setting up the discontinuous Galerkin method

The discontinuous Galerkin method can be generalised to a system in a straightforward manner. Each component of the approximate solution vector that we shall denote by $\mathbf{U}_h$ is defined locally in each cell as a polynomial of degree k, *i.e.* an element of $\mathbb{P}_k[x_\nu, x_{\nu+1}]$. Denoting by $\varphi_0, \ldots, \varphi_k$ a basis of $\mathbb{P}_k[x_\nu, x_{\nu+1}]$, the restriction of $\mathbf{U}_h$ to the cell $[x_\nu, x_{\nu+1}]$, that we shall denote by $\mathbf{U}_h^{\nu+\frac{1}{2}}$ can be expressed as

$$\mathbf{U}_h^{\nu+\frac{1}{2}}(t,x) = \sum_{j=0}^{k} \varphi_j(x) \mathbf{U}_j^{\nu+\frac{1}{2}}(t), \qquad \text{with } \mathbf{U}_j^{\nu+\frac{1}{2}}(t) = \begin{pmatrix} u_{j,1}^{\nu+\frac{1}{2}}(t) \\ \vdots \\ u_{j,d}^{\nu+\frac{1}{2}}(t) \end{pmatrix}.$$

The unknown function $\mathbf{U}_h(t,x)$ is completely determined by its components $\mathbf{U}_j^{\nu+\frac{1}{2}}(t)$. In order to determine those we proceed like in the scalar case and plug this expression in the equation, multiply by each of the basis functions and integrate the derivative term by part to let the flux through the interface appear:

$$\frac{d\mathbf{U}_j^{\nu+\frac{1}{2}}(t)}{dt} \int_{x_\nu}^{x_{\nu+1}} \varphi_i(x)\varphi_j(x)\,\mathrm{d}x - \int_{x_\nu}^{x_{\nu+1}} A\mathbf{U}_j^{\nu+\frac{1}{2}}(t)\varphi_i'(x)\varphi_j(x)\,\mathrm{d}x + \mathbf{G}_{\nu+1}\varphi_i(x_{\nu+1}) - \mathbf{G}_\nu \varphi_i(x_\nu) = 0.$$

The numerical flux $\mathbf{G}_\nu$ is a consistent approximation of the real flux $A\mathbf{U}(t,x_\nu)$ at the cell interface x_ν which is identical for both cells sharing x_ν.

Choice of the numerical flux: As in the scalar case, there are many possible choices of the numerical flux.

1. Centred flux: Use an average from the fluxes coming from both cells $\mathbf{G}_\nu = \frac{1}{2} A \left(\mathbf{U}_{\nu-\frac{1}{2}}(t,x_\nu) + \mathbf{U}_{\nu+\frac{1}{2}}(t,x_\nu) \right)$.

2. Solution of the Riemann problem corresponding to $\mathbf{U}_L = \mathbf{U}_{\nu-\frac{1}{2}}(t,x_\nu)$ and $\mathbf{U}_R = \mathbf{U}_{\nu+\frac{1}{2}}(t,x_\nu)$. This is the generalisation of the upwind flux to systems. Better stability properties but more diffusive.

3. Linear combination of the above to combine good stability properties of the latter and lower diffusion properties of former.

4. Generalised Lax-Friedrichs flux (also called Rusanov flux).

   $$\mathbf{G}_\nu = \frac{1}{2} \left[A \left(\mathbf{U}_{\nu-\frac{1}{2}}(t,x_\nu) + \mathbf{U}_{\nu+\frac{1}{2}}(t,x_\nu) \right) - \alpha_\nu \left(\mathbf{U}_{\nu+\frac{1}{2}}(t,x_\nu) - \mathbf{U}_{\nu-\frac{1}{2}}(t,x_\nu) \right) \right],$$

   with $\alpha_\nu = \max |\lambda_k|$, where the λ_k are the eigenvalues of A. This solver has good stability properties without needing to solve the Riemann problem. Especially interesting in the non linear case when the Riemann problem is hard to solve.

Chapter 4

Non linear conservation laws

4.1 Characteristics

We consider a generic scalar conservation law of the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \qquad \text{with the initial condition } u(0,x) = u_0(x).$$

Assuming $f \in \mathcal{C}^1(\mathbb{R})$ and denoting by $a(u) = f'(u)$, u also verifies

$$\frac{\partial u}{\partial t} + a(u)\frac{\partial u}{\partial x} = 0.$$

We now define the characteristic curves (or characteristics) associated to the conservation law the solution of the differential equation

$$\frac{dX}{dt} = a(u(t,X)).$$

Then using the chain rule we have

$$\frac{d}{dt}(u(t,X(t)) = \frac{\partial u}{\partial t}(t,X(t)) + \frac{dX}{dt}\frac{\partial u}{\partial x}(t,X(t)) = \frac{\partial u}{\partial t}(t,X(t)) + a(u(t,X(t))\frac{\partial u}{\partial x}(t,X(t)) = 0,$$

if u is a solution of the equation. From this we deduce that $u(t,X(t)) = u(0,X(0)) = u_0(X(0))$ which is independent of t. It follows that the characteristic curves are the straight lines in the (t,x) plane of equation:

$$X(t) = X(0) + a(u_0(X(0)))t.$$

And it follows that the solutions of the conservation law satisfy $u(t,X(t)) = u_0(X(0))$. This allows us to get the solution at a given point x and time t if the characteristic curve can be traced back in the (t,x) plane to the line $t = 0$. This is not always the case when f is non linear.

Remark 11 *In the linear case we have $f(u) = au$, then the characteristics are the parallel lines of slope a. They obey the equation $X(t) = X(0) + at$. So they never cross and so taking $X(t) = x$, we have $X(0) = x - at$, and we recover the classical solution of the linear advection equation*

$$u(t,x) = u_0(x - at).$$

Example. Burgers equation. It corresponds to $f(u) = \frac{1}{2}u^2$, so that $a(u) = f'(u) = u$. Then the characteristic $X(t;x)$ such that $X(0) = x$ satisfies $X(t;x) = x + tu_0(x)$. Hence if we consider $x_2 > x_1$, we have

$$X(t;x_2) - X(t;x_1) = x_2 - x_1 + t(u_0(x_2) - u_0(x_1)).$$

Then if u_0 is non decreasing we have $X(t;x_2) > X(t;x_1)$ for all positive times, but if u_0 is strictly decreasing then $X(t;x_2) = X(t;x_1)$ for $t_* = \frac{x_2 - x_1}{u_0(x_1) - u_0(x_2)}$. So the characteristics can cross and in this case the method of characteristics cannot be used to compute the solution which is then no more C^1.

4.2 Weak solutions

4.2.1 Definition

As we have seen in the case of Burgers equation, the characteristics associated to a non linear conservation law can cross, in which case the method of characteristics can non longer be used to find a solution. Actually when this happens the solution is no longer of class C^1 and can even become discontinuous. We thus need another form of the equation $\frac{\partial u}{\partial u} + \frac{\partial f(u)}{\partial x} = 0$ associated to the initial condition $u(0,x) = u_0(x)$ to make it well defined for such absolution.

Let us first recall the definition of a classical solution:

Definition 5 *Let $u_0 \in C^1(\mathbb{R})$, then u is called a classical solution if $u \in C^1(\mathbb{R}^+ \times \mathbb{R})$ and u satisfies*

$$\frac{\partial u}{\partial u} + \frac{\partial f(u)}{\partial x} = 0, \qquad \text{and } U(0,x) = u_0(x)$$

in the classical sense.

Now we can define the notion of weak solution:

Definition 6 *Let $u_0 \in L^\infty(\mathbb{R})$, then $u \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$ is called a weak solution of our scalar conservation law if u satisfies*

$$\int_0^T \int_{-\infty}^{+\infty} u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x}\,\mathrm{d}t\,\mathrm{d}x + \int_{-\infty}^{+\infty} u_0(x)\varphi(0,x)\,\mathrm{d}x = 0 \quad \forall \varphi \in C_c^1([0,T[\times\mathbb{R}.$$

Multiplying the equation $\frac{\partial u}{\partial u} + \frac{\partial f(u)}{\partial x} = 0$ by φ and integrating by parts, it is easy to verify that a classical C^1 solution is also a weak solution. So the notion of weak solutions generalises the notion of classical solutions.

4.2.2 The Rankine-Hugoniot condition

In most physical cases the solution is actually piecewise C^1 and there is only one (or a few) lines of discontinuities in the space-time plane. The Rankine-Hugoniot condition gives a constraint on the discontinuity along a line for the piecewise smooth solution to be a weak solution of the equation.

Theorem 5 *Assume the half space $\mathbb{R}^+ \times \mathbb{R}$ is split into two parts M_1 and M_2 by a smooth curve S parametrized by $(t, \sigma(t))$ with $\sigma \in C^1(\mathbb{R}^+)$. We also assume that $u \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$*

and that $u_1 = u_{|M_1} \in C^1(\bar{M}_1)$ *and* $u_2 = u_{|M_2} \in C^1(\bar{M}_2)$ *with* u_1 *and* u_2 *two classical solutions of our equation in respectively* M_1 *and* M_2*.*

Then u *is a weak solution if and only if*

$$[u_1(t,\sigma(t)) - u_2(t,\sigma(t))]\,\sigma'(t) = f(u_1(t,\sigma(t))) - f(u_2(t,\sigma(t))) \quad \forall t \in \mathbb{R}_+^*. \tag{4.1}$$

Relation (4.1) is called the Rankine-Hugoniot condition. It is often written in the simplified manner

$$(u_1 - u_2)s = f(u_1) - f(u_2),$$

where $s = \sigma'(t)$ is the propagation speed of the discontinuity.

Proof. Assume u is a weak solution of our equation. Then by definition

$$\int_0^T \int_{-\infty}^{+\infty} u\frac{\partial\varphi}{\partial t} + f(u)\frac{\partial\varphi}{\partial x}\,\mathrm{d}t\,\mathrm{d}x + \int_{-\infty}^{+\infty} u_0(x)\varphi(0,x)\,\mathrm{d}x = 0 \quad \forall\varphi \in C_c^1([0,T[\times\mathbb{R}.$$

Then we can split the first double integral into an integral on M_1 and an integral on M_2 and integrate these integrals by parts as u is C^1 on each of these domains. We then get using Green's divergence formula, with ν denoting the outward unit normal vector,

$$\int_\Omega \psi\nabla\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} = -\int_\Omega \mathbf{v}\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\int_{\partial\Omega}\mathbf{v}\cdot\nu\psi\,\mathrm{d}s$$

that

$$\int_{M_1} u\frac{\partial\varphi}{\partial t} + f(u)\frac{\partial\varphi}{\partial x}\,\mathrm{d}t\,\mathrm{d}x = -\int_{M_1}\left(\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x}\right)\varphi\,\mathrm{d}t\,\mathrm{d}x + \int_{\partial M_1}\begin{pmatrix}u\\ f(u)\end{pmatrix}\cdot\nu_1\varphi\,\mathrm{d}s$$

Because $u = u_1$ is a classical solution in M_1 the first term on the right hand side vanishes. The boundary term is composed of a part on the $t = 0$ line which cancels with the part of the integral on the initial condition which is in M_1 and a part that can be parametrized using σ. A tangent vector to the parametrized curve S is given by $(1, \sigma'(t))$. Hence the outward unit normal is given by

$$\nu_1 = \frac{1}{\sqrt{1+\sigma'(t)^2}}\begin{pmatrix}\sigma'(t)\\ -1\end{pmatrix}.$$

Recall that the integral over a curve parametrized by $\gamma : [a,b] \to \mathbb{R}^n$ is defined by

$$\int_S F\,\mathrm{d}s = \int_a^b F(\gamma(t))\|\dot{\gamma}(t))\|_2\,\mathrm{d}t.$$

Thus the part of the boundary integral corresponding to the integral on S which is parametrized by $\gamma(t) = (t, \sigma(t))$ becomes

$$\begin{aligned}\int_S \begin{pmatrix}u_1\\ f(u_1)\end{pmatrix}\cdot\nu_1\varphi\,\mathrm{d}s &= \int_0^T \frac{1}{\sqrt{1+\sigma'(t)^2}}\left[\sigma'(t)u_1(t,\sigma(t)) - f(u_1(t,\sigma(t)))\right]\varphi(t,\sigma(t))\\ &\qquad\qquad \sqrt{1+\sigma'(t)^2}\,\mathrm{d}t,\\ &= \int_0^T \left[\sigma'(t)u_1(t,\sigma(t)) - f(u_1(t,\sigma(t)))\right]\varphi(t,\sigma(t))\,\mathrm{d}t.\end{aligned}$$

In the same way for the part on M_2 for which the outward normal is $\nu_2 = -\nu_1$ we get

$$\int_S \begin{pmatrix} u_1 \\ f(u_1) \end{pmatrix} \cdot \nu_1 \varphi \, \mathrm{d}s = \int_0^T \left[-\sigma'(t) u_2(t, \sigma(t)) + f(u_2(t, \sigma(t))) \right] \varphi(t, \sigma(t)) \, \mathrm{d}t.$$

Adding the two pieces we find

$$\int_0^T \left[\sigma'(t)(u_1(t, \sigma(t)) - u_2(t, \sigma(t))) + f(u_2(t, \sigma(t))) - f(u_1(t, \sigma(t))) \right] \varphi(t, \sigma(t)) \, \mathrm{d}t = 0.$$

This is true for all C^1 functions φ, hence its factor in the integral vanishes. This corresponds to the Rankine-Hugoniot relation.

Conversely, the same calculation proves that if the Rankine-Hugoniot relation is satisfied then u is a weak solution. ■

Example: Consider the Burgers equation which corresponds to $f(u) = \frac{u^2}{2}$:

$$\frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial u^2}{\partial x} = 0,$$

with the initial condition

$$u_0(x) = \left| \begin{array}{lll} 0 & \text{if} & x < 0, \\ 1 & \text{if} & x \geq 0. \end{array} \right.$$

1. Using the Rankine-Hugoniot relation let us see under what condition a piecewise constant solution can be a weak solution of this problem. Because the characteristics of the Burgers equation are straight lines, it is natural to look for piecewise smooth solutions which are separated by a straight line in the $t - x$ plane. A straight line can be parametrized by $(t, \alpha t)$. Then the candidate piecewise solution writes

$$u(t, x) = \left| \begin{array}{lll} 0 & \text{if} & x < \alpha t, \\ 1 & \text{if} & x \geq \alpha t. \end{array} \right.$$

   As $\sigma(t) = \alpha t$, we have $s = \sigma'(t) = \alpha$ and so the Rankine-Hugoniot condition $(u_1 - u_2)s = f(u_1) - f(u_2)$ becomes

$$(0 - 1)\alpha = \frac{1}{2}(0 - 1)$$

   and thus the only weak solution consisting of two constant states is u corresponding to $\alpha = \frac{1}{2}$. Such a discontinuous solution is called a *shock*.

2. Let us now define

$$u(t, x) = \left| \begin{array}{lll} 0 & \text{if} & x < 0, \\ x/t & \text{if} & 0 \leq x < t, \\ 1 & \text{if} & x \geq t. \end{array} \right.$$

   The two constant values are obviously classical solutions in their part of the domain. On the other hand $u(t, x) = x/t$ verifies $\frac{\partial u}{\partial t}(t, x) = -\frac{x}{t^2}$ and $\frac{\partial u^2}{\partial x}(t, x) = \frac{2x}{t^2}$ so that

$$\frac{\partial u}{\partial t}(t, x) + \frac{1}{2} \frac{\partial u^2}{\partial x}(t, x) = -\frac{x}{t^2} + \frac{1}{2} \frac{2x}{t^2} = 0$$

   and $u(t, x) = x/t$ is also a classical solution on its domain of definition. It is straightforward to verify that this solution is continuous on the lines $x = 0$ and

$x = t$, so that the Rankine-Hugoniot conditions are automatically verified (no jump) but not C^1. This solution which goes continuously from one constant state to the other, but is not C^1 on two lines is called a *rarefaction wave*.

So we proved that this is also a weak solution of the Burgers equation, which means that the equation we are considering in this example has at least two weak solutions. In practice it is possible to construct other weak solutions with several constant states separated by a line of discontinuity (see [9] for examples).

4.2.3 Entropy solution

The last example shows us that the uniqueness of a weak solution is not guaranteed. However the physical solution is unique. This means that there is a piece missing in our theory that will enable us to characterise the physically correct weak solution. The idea that will lead us there is that an exact conservation law is unlikely to be physical. There is always a small amount of dissipation. Mathematically this can be modelled by adding a small diffusion term to our conservation law: for a small positive ϵ we consider

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} - \epsilon \frac{\partial^2 u}{\partial x^2} = 0. \tag{4.2}$$

Associated to a smooth initial condition this equation has a unique smooth solution. And the physically correct solution of our conservation law, will be the unique limit of this solution (which depends on ϵ) when $\epsilon \to 0$. This unique solution can be characterised using the notion of entropy.

Definition 7 *Let $U, F \in C^2(\mathbb{R})$ such that*

(i) U is strictly convex,

(ii) $F' = U' f'$

Then U is called an entropy and F an entropy flux associated to the conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0.$$

Theorem 6 *Assume the conservation law $\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$ can be associated to an entropy U with entropy flux F. Then if $(u_\epsilon)_\epsilon$ is a sequence of smooth solutions of (4.2) such that $\|u_\epsilon\|_{L^\infty}$ is bounded uniformly in ϵ and $u_\epsilon \to u$ almost everywhere then u is a weak solution of the conservation law and verifies the entropy condition*

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \leq 0 \tag{4.3}$$

in the weak sense, i.e.

$$\int_0^T \int_{-\infty}^{+\infty} (U(u)\frac{\partial \varphi}{\partial t} + F(u)\frac{\partial \varphi}{\partial x})\,\mathrm{d}t\,\mathrm{d}x \geq 0 \quad \forall \varphi \in C_c^1(]0,T[\times\mathbb{R}), \varphi \geq 0.$$

Definition 8 *A weak solution that verifies the entropy condition (4.3) is called an entropy solution.*

It has been proven by Kruzhkov (see [9] and references therein) that for a bounded initial condition u_0 and a smooth flux $f(u)$, there exists a unique entropy solution of the scalar conservation law.

In the case of a strictly convex flux ($f''(u) > 0$) which is verified by the Burgers equation, there is a simple characterisation of the entropy condition for a shock, which is called the Lax entropy condition:

$$f'(u_L) > s > f'(u_R), \quad \text{with } s = \frac{f(u_R) - f(u_L)}{u_R - u_L}.$$

This is the Rankine-Hugoniot setting and s is the shock speed. This tells us that the shock is the entropy solution provided the characteristics lines cross at the shock line $x = st$. Else the entropy solution is a rarefaction wave.

Remark 12 *Because f is strictly convex f' is increasing and the entropy condition can be satisfied only if $u_L > u_R$ and in this case because of the strict convexity, s which is the slope of the line joining $(u_L, f(u_L))$ and $(u_R, f(u_R))$ on the graph of f lies necessarily between $f'(u_R)$ and $f'(u_L)$.*

Example: Consider again the Burgers equation $\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0$ with piecewise constant initial conditions $u_0(x) = u_L$ for $x < 0$ and $u_0(x) = u_R$ for $x \geq 0$. We compute previously the characteristics for Burgers equation which are given by $X(t; x) = x + tu_0(x)$. So in our case we have two families of characteristics, those issued from the negative values of x at $t = 0$ which become $X(t; x) = x + tu_L$ and those issued from the positive values of x at $t = 0$ which become $X(t; x) = x + tu_R$. The flux of the Burgers equation $f(u) = \frac{1}{2}u^2$ is strictly convex, so the Lax criterion applies and the Lax entropy condition can be satisfied only if $u_L > u_R$ which is the case when the characteristics cross.

Remark 13 *In the case of a linear flux $f(u) = au$ for some constant a. If U is an entropy, the associated entropy flux is $F = au$ and we have, as u satisfies $\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$,*

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} = \frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = U'(t)(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}) = 0.$$

So the entropy condition is always verified in this case.

This is a simple case, which also can appear for some of the waves in non linear systems. In this case two different values are just propagated side by side without interfering. This case is called a contact discontinuity.

We will restrict in this lecture to the two simplest cases of scalar conservation laws, namely the cases of a strictly convex flux and of a linear flux. These often occur in applications, but not always. A more complex study is needed in other cases. We refer to the book of Leveque [11] for details.

4.3 The Riemann problem

Let us here compute the exact entropy solution of the Riemann problem for a strictly convex flux *i.e.* $f''(u) > 0$.

The Riemann problem for the 1D scalar non linear conservation laws reads

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0,$$

$$u_0(x) = \left| \begin{array}{lll} u_L & \text{if} & x < 0, \\ u_R & \text{if} & x \geq 0. \end{array} \right.$$

Case 1: $u_L < u_R$**.** We saw previously that in this case the shock does not satisfy the Lax entropy condition and is not entropic. We thus look for a rarefaction wave which is continuous across each of the characteristics issued from the left and right of 0. Between these two characteristics, we look for an explicit solution of the form $u(t,x) = v(x/t)$. Plugging this function into our equation, we get

$$-\frac{x}{t^2}v'(\frac{x}{t}) + f'(v(\frac{x}{t}))\frac{1}{t}v'(\frac{x}{t}) = 0.$$

Setting $\xi = x/t$ this becomes

$$(f'(v(\xi)) - \xi)v'(\xi) = 0.$$

So we get two kinds of solution, either $v(\xi) = C$ (C constant), or v such that $f'(v(\xi)) - \xi = 0$. The constant solution is not possible as it would yield non entropic shocks. As $f'' > 0$, f' is strictly increasing and thus invertible. Hence the solution can be expressed as $v(\xi) = (f')^{-1}(\xi)$.

Finally the entropy solution of the Riemann problem in this case is the rarefaction wave defined by

$$u(t,x) = \left| \begin{array}{ccc} u_L & \text{if} & x < f'(u_L)t, \\ (f')^{-1}(x/t) & \text{if} & f'(u_L)t \leq x < f'(u_R)t, \\ u_R & \text{if} & x \geq f'(u_R)t. \end{array} \right.$$

Case 2: $u_L > u_R$**.** In this case we saw that the shock separating the two constant states u_L and u_R and propagating at the speed defined by the Rankine-Hugoniot condition

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R}$$

is entropic and so is the unique physical solution. This is defined by

$$u(t,x) = \left| \begin{array}{lll} u_L & \text{if} & x < st, \\ u_R & \text{if} & x \geq st. \end{array} \right.$$

4.4 Numerical methods

4.4.1 The Godunov method

The idea of the Godunov method is that if one is looking for piecewise constant solutions then as the solution is propagating at a finite speed which is known, the solution can be computed exactly by solving a Riemann problem for each cell interface. If $s = \max |f'(u)|$ denotes the fastest possible wave then the neighbouring Riemann problems do note interact provided the time step Δt verifies $s\Delta t \leq \frac{1}{2}\Delta x$. In principle the solution at time t_{n+1} could be computed by integrating over the solutions of all the Riemann problems at this time to find the new constant value on each cell by averaging. However, this could

be quite complicated. The Godunov method can be expressed in a much simpler way by integrating the starting equation in space and time over one cell and one time step:

$$\int_{t_n}^{t_{n+1}} \int_{x_i}^{x_{i+1}} \left(\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} \right) \mathrm{d}x \, \mathrm{d}t = 0.$$

Denoting by $u_{i+\frac{1}{2}}^n = \frac{1}{\Delta x} \int_{x_i}^{x_{i+1}} u(t_n, x) \, \mathrm{d}x$ the average of u over a cell at time $t_n = n\Delta t$, this yields

$$\Delta x (u_{i+\frac{1}{2}}^{n+1} - u_{i+\frac{1}{2}}^n) + \int_{t_n}^{t_{n+1}} (f(u(t, x_{i+1})) - f(u(t, x_i))) \, \mathrm{d}t = 0.$$

Now as u is assumed to be constant on each cell and equal to its average, the value of the flux $f(u(t, x_i))$ can be exactly computed by solving the Riemann problem between the left value $u_{i-\frac{1}{2}}^n$ and the right value $u_{i+\frac{1}{2}}^n$ at the interface x_i. Note that we only need the solution of the Riemann problem exactly on the interface and this solution does not depend on time as long as there is no interaction with the neighbouring problem on the interface, which cannot occur if $s\Delta t \leq \Delta x$. Here because we are only interested in the solution directly at the interface we gain a factor $\frac{1}{2}$ as we do not care about other interactions.

As we saw previously, assuming that the flux $f(u)$ is strictly convex, the entropy solution of the Riemann problem can be either a shock or a rarefaction wave, but $u(t, x_i)$ will take a value different from $u_{i-\frac{1}{2}}^n$ or $u_{i+\frac{1}{2}}^n$ only if the rarefaction wave has an interaction with $x = x_i$ which is the case only if $f'(u_L) < 0$ and $f'(u_R) > 0$. Let us clearly distinguish the cases and express the solution of the local Riemann problem at the interface $x = x_i$. Recall that we consider the local Riemann problem with the initial condition (at time t_n)

$$u(t_n, x) = \begin{cases} u_{i-\frac{1}{2}}^n & \text{if} \quad x < x_i, \\ u_{i+\frac{1}{2}}^n & \text{if} \quad x \geq x_i. \end{cases}$$

If $u_{i-\frac{1}{2}}^n < u_{i+\frac{1}{2}}^n$ the entropy solution is a rarefaction wave so that

$$u(t, x_i) = \begin{cases} u_{i-\frac{1}{2}}^n & \text{if} \quad f'(u_{i-\frac{1}{2}}^n) > 0, \\ (f')^{-1}(0) & \text{if} \quad f'(u_{i-\frac{1}{2}}^n) \leq 0 \leq f'(u_{i+\frac{1}{2}}^n), \\ u_{i+\frac{1}{2}}^n & \text{if} \quad f'(u_{i+\frac{1}{2}}^n) < 0. \end{cases}$$

If $u_{i-\frac{1}{2}}^n > u_{i+\frac{1}{2}}^n$ the entropy solution is a shock wave so that

$$u(t, x_i) = \begin{cases} u_{i-\frac{1}{2}}^n & \text{if} \quad s > 0, \\ u_{i+\frac{1}{2}}^n & \text{if} \quad s < 0, \end{cases}$$

where s is the shock speed given by the Rankine-Hugoniot condition: $s = \frac{f(u_L) - f(u_R)}{u_L - u_R}$.

Using the exact solution of the Riemann problem, the numerical flux $g_i^n = g(u_{i-\frac{1}{2}}^n, u_{i+\frac{1}{2}}^n)$ is taken to be the flux associated to the exact solution of the Riemann problem, *i.e.* $f(u(t, x_i))$ for $x = x_i$ and $t > t_n$. Noticing that for a strictly convex flux, the minimum of f is reached at the points $(f')^{-1}(0)$ and going through the different cases we find that

$$g_i^n = \begin{cases} \min\limits_{u \in [u_{i-\frac{1}{2}}, u_{i+\frac{1}{2}}]} f(u) & \text{if} \quad u_{i-\frac{1}{2}} < u_{i+\frac{1}{2}}, \\ \max\limits_{u \in [u_{i-\frac{1}{2}}, u_{i+\frac{1}{2}}]} f(u) & \text{if} \quad u_{i-\frac{1}{2}} > u_{i+\frac{1}{2}}. \end{cases}$$

With this, the flux at the interface can be computed and the Godunov scheme is well defined (here in the case of a strictly convex flux). In the case of a linear flux, the Godunov scheme amounts to the first order upwind scheme. For other smooth fluxes, the Riemann problem can generally also be solved even though it is more complicated.

4.4.2 Approximate Riemann solvers

Computing the exact solution of the local Riemann problems is in principle the best method. However it can be complicated or numerically expensive. Moreover, as there are other sources of numerical errors anyway, it might be numerically as good or almost as good to replace the solution of the exact Riemann problem by some approximation.

1. The Rusanov flux (also called Local Lax-Friedrichs flux). The idea behind this flux, instead of approximating the exact Riemann solver, is to recall that the entropy solution is the limit of viscous solutions and to take a centred flux to which some viscosity (with the right sign) is added. This flux is given by

$$g_i^n = g(u^n_{i-\frac{1}{2}}, u^n_{i+\frac{1}{2}}) = \frac{1}{2}\left[f(u_{i-\frac{1}{2}}) + f(u_{i+\frac{1}{2}}) - \max_{u\in[u_{i-\frac{1}{2}},u_{i+\frac{1}{2}}]} |f'(u)|(u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}})\right].$$

   Taking the viscosity parameter as the largest local wave speed guarantees the stability of the scheme. On the other hand taking the largest local wave speed (instead of globally) adds less viscosity in regions where the solution is smooth and $u_{i-\frac{1}{2}}$ is close to $u_{i+\frac{1}{2}}$.

2. The Roe flux (also called Murman or Murman-Roe in the scalar case): the idea here is to linearise the flux $f(u)$ around the cell interface and then use the upwind flux. Assuming two constant values u_L and u_R on each side, this amounts to replacing $f(u)$ by $a(u_L, u_R)u$ with a well chosen velocity $a(u_L, u_R)$ that will enable us to get a flux which is close to the flux given by the solution of the exact Riemann problem. Looking at the solution of the exact Riemann problem an approximation that looks interesting for both rarefaction wave and shocks is to take the Rankine-Hugoniot velocity

$$a(u_L, u_R) = \frac{f(u_L) - f(u_R)}{u_L - u_R},$$

   if there is a discontinuity *i.e.* $u_L \neq u_R$, and simply $a(u_L, u_R) = f'(u_L) = f'(u_R)$ if $u_L = u_R$. Indeed if u_L and u_R are close, which will be the case in smooth regions $a(u_L, u_R)$ defines a good approximation of both $f'(u_L)$ and $f'(u_R)$. Moreover if both $f'(u_L)$ and $f'(u_R)$ have the same sign $a(u_L, u_R)$ will also have this sign, so that the same value of the numerical flux will be obtained as in the exact Riemann problem was solved. The only case where this does not yield the same solution as the exact Riemann solver is the case of a rarefaction wave with $f'(u_L)$ and $f'(u_R)$ of different signs. In this case, it is possible that the solution obtained by the scheme is not the correct entropy solution, the non entropic shock being approximated instead of the rarefaction wave. So a fix, known as entropy fix, is needed. For this let us express the Murman-Roe flux as a centred flux with some added viscosity.

The Murman-Roe flux is an upwind flux for the linearised problem we just introduced. The case $u_L = u_R$ being trivial, we consider only $u_L \neq u_R$. Then

$$g(u_L, u_R) = \left| \begin{array}{ll} f(u_L) & \text{if} \quad a(u_L, u_R) > 0, \\ f(u_R) & \text{if} \quad a(u_L, u_R) \leq 0. \end{array} \right.$$

If $a(u_L, u_R) = \frac{f(u_L)-f(u_R)}{u_L-u_R} > 0$ and $u_L \neq u_R$ we have

$$f(u_L) = \frac{1}{2}\left(f(u_L) + f(u_R) - \frac{f(u_R) - f(u_L)}{u_R - u_L}(u_R - u_L) \right),$$

and if $a(u_L, u_R) = \frac{f(u_L)-f(u_R)}{u_L-u_R} < 0$ and $u_L \neq u_R$ we have

$$f(u_R) = \frac{1}{2}\left(f(u_L) + f(u_R) + \frac{f(u_R) - f(u_L)}{u_R - u_L}(u_R - u_L) \right),$$

so that we can define the numerical flux in all cases by

$$g(u_L, u_R) = \frac{1}{2}\left(f(u_L) + f(u_R) - |a(u_L, u_R)|(u_R - u_L) \right).$$

Here we also see that the numerical viscosity vanishes when $a(u_L, u_R) \approx 0$ which can happen close to the minimum of the convex function f. Then a non entropic shock might be selected by the scheme. A simple fix, introduced by Harten, consists in smoothing the graph of the absolute value close to 0 (see [11] for details). This consists in replacing the absolute value in the formula defining the flux by

$$\phi(\lambda) = \left| \begin{array}{ll} |\lambda| & |\lambda| \geq \epsilon, \\ (\lambda^2 + \epsilon^2)/(2\epsilon) & |\lambda| < \epsilon. \end{array} \right.$$

This ensures that $\phi(\lambda) \geq \epsilon$ and that there is always some dissipation. This works and yields the correct entropy solution provided ϵ is well tuned to the problem at hand.

4.4.3 Higher order methods

Higher order methods can be designed using the Finite Volume or the Discontinuous Galerkin methodology. In both cases what needs to be defined is the numerical flux constructed from the two values u_L and u_R on each side of the interface. The same numerical fluxes as previously can be used in this case. However for the scheme to be stable close to discontinuities a limiting procedure needs to be introduced so that the scheme falls back to first order in the cells where a discontinuity is detected. There is a vast literature on limiters that we shall not detail here.

4.4.4 Strong stability preserving (SSP) Runge-Kutta methods.

Fluxes are generally constructed so that the associated scheme has some stability properties when used with a first order explicit Euler time solver. When higher order methods in x are used, it makes sense to use also higher order methods in time. A specific class of Runge-Kutta methods has been developed to keep the strong stability properties of the explicit Euler solver at each stage (See [12]).

4.5 Nonlinear systems of conservation laws

Going from the scalar case to systems in the non linear case, is similar to what is done in the linear case. The hyperbolicity of the system is essential so that the system can be locally diagonalised and the eigenvalues explicitely used in the definition of the flux.

The derivation of a Finite Volume or Discontinuous Galerkin scheme can be done component by component and so reduces to the scalar case except for the definition of the numerical flux which in general mixes the different components and needs to be specific to the system at hand. We shall restrict in this lecture to the introduction of two of the most used numerical fluxes, namely the Rusanov (or local Lax-Friedrichs) flux and the Roe flux.

4.5.1 The Rusanov flux

As in the scalar case, the main idea here is to use a centred flux to which just enough dissipation is added to ensure stability in all cases. In the scalar case the needed viscosity was given by the largest local wave speed. A system of n components corresponds to the superposition of n waves the local speed of each being given by the corresponding eigenvalue. So taking the viscosity coefficient in the flux as the maximum over all eigenvalues should do the job. This yields the Rusanov flux for systems, which is the simplest stable flux. It is defined for a linear system of the form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0,$$

as

$$\mathbf{G}(\mathbf{U}_L, \mathbf{U}_R) = \frac{1}{2}\left(\mathbf{F}(\mathbf{U}_L) + \mathbf{F}(\mathbf{U}_R) - \max_{U\in[U_L,U_R]} |\lambda(\mathbf{F}'(\mathbf{U}))|(\mathbf{U}_R - \mathbf{U}_L)\right),$$

where $\max_{U\in[U_L,U_R]} |\lambda(\mathbf{F}'(\mathbf{U}))|$ denotes the maximum modulus of the eigenvalues of the Jacobian matrix $\mathbf{F}'(\mathbf{U})$.

4.5.2 The Roe flux

Roe's method consists in locally linearising the non linear flux with a well chosen procedure. The linearised matrix between two constant states $\mathbf{U}_L$ and $\mathbf{U}_R$ is denoted by $A(\mathbf{U}_L, \mathbf{U}_R)$ and constructed such that the following properties are verified:

- $\mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = A(\mathbf{U}_L, \mathbf{U}_R)(\mathbf{U}_R - \mathbf{U}_L)$.
- $A(\mathbf{U}, \mathbf{U}) = \mathbf{F}'(\mathbf{U})$.
- $A(\mathbf{U}_L, \mathbf{U}_R)$ is diagonalisable, has real eigenvalues and a complete system of eigenvectors.

Such a matrix is not always easy to find, but there are procedures, described in [11] for example, to construct them. Moreover classical Roe matrices are known for the most usual systems [11].

Once the Roe matrix is defined, the flux can be computed by solving the corresponding linear Riemann problem that we treated in Chapter 3. Let us rewrite the formula, so that we can also include the entropy fix, which as in the scalar case is needed for non linear systems to make sure that the scheme always converges to the correct entropy solution.

In the case of linear systems, the flux was defined as $AU(t,0) = A_+U_L + A_-U_R$. Defining the absolute value of a matrix as $|A| = A_+ - A_-$, the flux can also be expressed as

$$AU(t,0) = \frac{1}{2}\left(AU_L + AU_R - |A|(U_R - U_L)\right).$$

Using the properties of the Roe matrix in the non linear case, the same expression will be used to define the Roe flux:

$$\mathbf{G}(\mathbf{U}_L, \mathbf{U}_R) = \frac{1}{2}\left(A(\mathbf{U}_L, \mathbf{U}_R)\mathbf{U}_L + A(\mathbf{U}_L, \mathbf{U}_R)\mathbf{U}_R - |A(\mathbf{U}_L, \mathbf{U}_R)|(\mathbf{U}_R - \mathbf{U}_L)\right).$$

The same entropy fix consisting in replacing the absolute value by the function ϕ which is bounded away from 0 can be applied in this formula.

Bibliography

[1] D. N. Arnold, R. S. Falk, R. Winther. Finite element exterior calculus, homological techniques, and applications. Acta Numer. 15 (2006) 1–155.

[2] D. N. Arnold, R. S. Falk, R. Winther. Finite element exterior calculus: From Hodge theory to numerical stability Bull. AMS 47 (April 2010) 281–354.

[3] J.-P. Berrut, L.N. Trefethen. Barycentric Lagrange Interpolation, SIAM Rev. 46 (3), pp. 501–517.

[4] G. Cohen. *Higher-Order Numerical Methods for Transient Wave equation*, Springer-Verlag, 2001.

[5] G. Cohen, P. Monk. Efficient Edge Finite Element Schemes in Computational Electromagnetism, Proc. of the 3rd Conf. on Mathematical and Numerical Aspects of Wave Propagation Phenomena, SIAM, april 1995.

[6] D. A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*, vol. 69 SMAI Mathématiques et Applications, Springer, 2012.

[7] A. Ern, J.-L. Guermond. *Theory and Practice of Finite Elements*, Springer 2004.

[8] M. Gerritsma, Edge functions for spectral element methods. *In: Spectral and High Order Methods for Partial Differential Equations.* Springer Berlin Heidelberg, 2011. S. 199-207.

[9] E. Godlewski and P.A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, 1996.

[10] R. Hiptmair. Finite elements in computational electromagnetism, Acta Numerica (2002), pp. 237-339.

[11] R. J. Leveque. Finite Volume Methods for Hyperbolic Problems, Cambridge Texts in Applied Mathematics, 2002.

[12] J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin methods*, Springer, 2008.

[13] J.-C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. Numer. Math. 35 (1980), pp. 315–341.

[14] J.-C. Nédélec. A new family of mixed finite elements in R^3. Numer. Math. 50 (1986), no. 1, 57–81.

[15] Chi-Wang Shu. High Order Weighted Essentially Nonoscillatory Schemes for Convection Dominated Problems. SIAM Rev. 51 (1), pp. 82–126.