SCIENCES
SORBONNE
UNIVERSITÉ

Numerical fluid plasmas

5PYP05

roch smets — 2024

Contents

Preamble

This course is an introduction to numerical methods dedicated to the fluid treatment of plasmas. Most of the basis and general ideas are then discussed, but not necessarily in deepth.

- For the student who will continue with numerical simulations during his PhD, this course presents the minimum concepts that you will have to deepen.
- For the student who will not do by himself numerical simulations during his PhD, this course should be enough to understand the frame, constraints and flaws of numerical schemes and their associated numerical solutions.

Depending on the study case, the equations to be numerically solved will be scalar or vectorial. The generalization from scalar to vectorial form is not always trivial...

There is, as much as needed, **notation** (aside from **remark**, **example**, ...) in order to enlight the proper way to translate a concept in equation. From [Laney, 1998] (see page 75) : "the main obstacle is not finding the solution, but finding the proper notation for expressing the solution".

I also have to specify that there is a very wide panel of books, reports, articles, courses, notes... on this topic. I off course used many of them ; I tried to provide the citation whenever possible, but obviously, many of them are certainly lacking ; mainly because they do not have any reference number (like a `doi`) nor permanent `URL` where to be downloaded.

- The book by [Morton and Mayers, 1994] is a very nice introduction to all kinds of PDEs. It is mainly dedicated to finite difference methods, but introduces most of the concepts very clearly... and this is a very concise book.
- The book by [Laney, 1998] is very complete for the hyperbolic equations. He treats many aspects of the very last solvers "on the market". The case of neutral fluids (mainly with the associated Euler equations) is the only case discussed, excluding the case of plasmas.
- The book by [Godlewski and Raviart, 1996] is also very exhaustive. As it is written by mathematicians (Pierre Arnaud Raviart was a key french researcher in this field) physical implications are not deeply discussed, but the mathematical approach is extensively investigated.
- The book by [Toro, 2009] is specifically dedicated to the numerical treatment of the Riemann problem.

These references make clear that thos course is totally dedicated to the treatment of hyperbolic equations. Any dissipative term like diffusion is hence no treated.

— 1 —

Basics in fluid plasma equations

1.1 Classification

Let's consider the partial differential equation (PDE) on $u(x, y)$ of the form

$$a\partial_{x^2}^2 u + b\partial_x\partial_y u + c\partial_{y^2}^2 u + d\partial_x u + e\partial_y u + fu = 0 \tag{1.1}$$

Notation 1. *In this course, the index notation x, y, z and t refers to the derivative with respect to x, y and z spatial coordinate and time coordinate, respectively.*

Eq. (1.1) is

- hyperbolic if and only if $b^2 - 4ac > 0$
- parabolic if and only if $b^2 - 4ac = 0$
- elliptic if and only if $b^2 - 4ac < 0$

Remark 1. *The terminology used for this definition is derived from the conic (of the plan) classification.*

Remark 2. *If any of the coefficients $a, \dots f$ depends on x and/or y, the PDE is then said to be* **local**.

Remark 3. *The kind of a PDE does not depend on the base.*

1.1.1 Elliptic equations

Elliptic equations are usually governing stationnary and bounded problems, defined on a domain Ω for which the boundary conditions (defined on $\partial\Omega$) are usually of Dirichlet- or Neumann-type. The very classical elliptic equation is the Poisson equation

$$\begin{cases} \Delta u &= -f \qquad \text{on } \Omega \\ u &= u_0 \qquad \text{on } \partial\Omega \end{cases} \tag{1.2}$$

Example 1. *The Maxwell-Gauss equation governing the scalar potential (associated to the electrostatic part of the electric field) in a plasma is an elliptic equation.*

1.1.2 Parabolic equations

Parabolic equations are governing the time evolution of systems where diffusion (or any kind of dissipation) processes are at play. These problems are usually defined in a bounded domain Ω where the conditions on $\partial\Omega$ are Dirichlet- or Neumann-type (and eventually non-stationnary), but also involve some initial conditions.

Example 2. *The heat equation with Dirichlet conditions is the most classical case of parabolic equation. It has an analytic solution which depends on Green functions, involving both initial and boundary conditions.*

$$\left\{\begin{array}{lcll} \partial_t T & = & D\partial^2_{x^2} T & \text{on } \Omega \text{ with } D > 0 \\ T & = & T_0 & \text{on } \partial\Omega \text{ and } T(x,0) = f(x) \text{ on } \Omega \end{array}\right. \tag{1.3}$$

1.1.3 Hyperbolic equations

These equations are associated to dissipation-less wave propagation (otherwise, any dissipation would involve a second order term so the equation would then be parabolic). For a constant value of the phase speed (like in the Maxwell equations describing the propagation of a light wave) the hyperbolic equation is linear ; in the other cases (like the Euler equations governing an inviscid hydrodynamic flow), the equation is non-linear.

A first form of hyperbolic equation is

$$\partial^2_{t^2} u - A^2 \partial^2_{x^2} u = 0 \tag{1.4}$$

and is the equation describing a fluctuation $u(x,t)$ moving at speed $+A$ or $-A$.

Example 3. *We indifferently speak of "fluctuations" or "waves", as the former ones can be Fourier decomposed as a linear combination of the latter ones.*

The other form of hyperbolic equation

$$\partial_t u + A\partial_x u = 0 \tag{1.5}$$

is the one associated to fluctuations only moving at speed $+A$.

Remark 4. *In Eq. (1.5), the speed A can be positive or negative (whatever it is a constant or a function).*

Property 1. *In the multidimensional case, that is involving $n > 1$ spatial coordinates, $\mathbf{A}$ turns to be a $n \times n$ matrix. In such a case, the equation is hyperbolic if the matrix $\mathbf{A}$ is diagonalizable.*

1.2 Hyperbolic PDE

While the fluid treatment of a plasma generally involves both advection and diffusion processes, the time integration of these equations oftenly involves a separations of these processes.

Any dissipative process has a regularizing effect on the system, and is associated to a parabolic PDE. The numerical solution of these problems are not treated in this course.

In treating advection problems, the numerical treatement of front steepening and shock formation, as well as rarefaction waves are very challenging. We then concentrate on these problems in this course.

In tis course, we consider the second form of hyperbolic equation given by Eq. (1.5).

1.2.1 Caracteristics

In order to introduce the total derivative of u, we need to specify a path $X(t)$ so that this total derivative writes

$$\mathrm{d}_t u|_X = \mathrm{d}_t u(X(t), t) = \partial_t u + \partial_x u \times \mathrm{d}_t x|_X \tag{1.6}$$

As a consequence, along $X(t)$, u is a fonction of t only. Hence, defining $X(t)$ as the solution of $\mathrm{d}_t x = A$, Eq. (1.5) writes

$$\partial_t u + A\partial_x u = 0 = \mathrm{d}_t u|_X \tag{1.7}$$

As a consequence, the solution of Eq. (1.5) can be reduced to the solution of the Ordinary Differential Equation (ODE) $\mathrm{d}_t u|_X = 0$. The curves solution of this ODE (each of them depending on the initial condition $x(t = 0) = x_0$ for the solutions of the ODE) are called the **caracteristics** of the hyperbolic equation. The promising method of caracteristics is then essentially a reformulation of the equation using a change of variable. This approach then leads to the set of coupled differential equations

$$\left\{ \begin{array}{rcl} \mathrm{d}_t x|_X & = & A \\ \mathrm{d}_t u|_X & = & 0 \end{array} \right. \tag{1.8}$$

Two difficulties then arises from this formulation

- this set of equation is "simple" if A is a constant or a function of x only. But if A is a function depending on t or u, this set is coupled and can generally not be decoupled in a straightforward way
- singularities can (and will) arise when caracteristics will cross ; this is associated to the steppening of $u(x, t)$ (like in shocks or discontinuities) in non-linear cases where dissipations effects are not strong enough to control such steppening

In the simple case of a constant speed A, the first caracteristic equation is $\mathrm{d}_t x = A$ which solution is then $X(t) = X_0 + At$ for $X(0) = X_0$. With the initial condition $u(x, 0) = u_0(x)$, the general solution is then $u(x, t) = u_0(x - At)$ and the caracteristics are straight lines of slope A^{-1} in the (x, t) plan.

Remark 5. *When the uniqueness of the solution is not guaranteed, a method (we should better say a "patch") consist in introducing a small but finite viscous term which origin is not physical, but strictly numerical, in order to preserve the regularity of the solution, leading to the equation* $\partial_t u + A\partial_x u = \varepsilon\Delta u$. *Such an approach is discussed in chapter 2.*

While there is a large panel of situations, we can roughly classify the kind of solutions in three categories that are illustrated on Fig. 1.1 :

- Progressive waves, illustrated in the left panel of Fig. 1.1. These waves advect the structures (with distorsion if A is not uniform) at speed A
- Expansion waves or rarefaction waves, illustrated in the middle panel of Fig. 1.1. These waves are expanding meaning that the distance between their head and tail is increasing with time.
- Shock waves and contact discontinuities, illustrated in the right panel of Fig. 1.1. These waves result from the steppening of a structure until infinite derivative, associated to the crossing of caracteristics.

Property 2. *For a shock wave travelling from left to right, pressure, density and wave speed are higher on the left side of the shock than on the right side.*

Definition 1. *A* **contact discontinuity** *is a structure where the total pressure are continuous, while the density and the temperature are discontinuous. Roughly speaking, a contact discontinuity separates two regions with different properties.*

Property 3. *The* **entropy** *increases across a* **shock** *while it is conserved across a* **contact discontinuity**.

Remark 6. *In a contact discontinuity, the discontinuity is co-moving with the fluid. The stationarity of this structure in the fluid frame is coherent with the pressure balance condition.*

Remark 7. *In a plasma, the magnetic field and the associated magnetic pressure modify this picture so that the total pressure includes the magnetic pressure. In the magnetized case, the discontinuity is a* **rotational dicccontinuity** *for which the normal flow velocity can be discontinuous. The contact discontinuity is hence the special case for which the normal component of the flow velocity is null (in the frame of the discontinuity).*

It is clear that the set of curves (one per initial condition at $t = 0$) displayed in Fig. 1.1 can be made for each variables of the problem (n, v and e for the one dimensional Euler problem). It is not a rule that at a given (x, t) point, all the quantities of the problem should behave in the same way (progressive, rarefaction or shock wave).

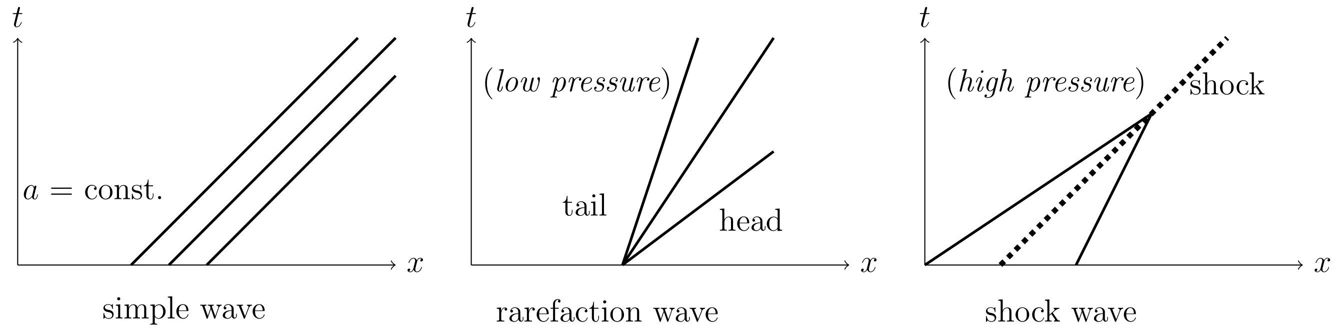

Figure 1.1: Illustration of caracteristics for a progressive wave (left panel) a rarefaction wave (middle panel) and a shock wave (right panel).

Remark 8. *The term "rarefaction" can clearly be understood when talking about the density, but apply for all the variables of the problem*

Remark 9. *In many situations, compressions and rarefactions can coexist aside one from the other, evolving, growing, shrinking, collapsing, splitting...*

Remark 10. *Shock creation is not always the fatal ending of a front stepening.*

In both the middle and right panel of Fig. 1.1, (rarefaction and compression, respectively), the slopes of the caracteristics are all positive. A singular situation can appear in regions where both $A > 0$ and $A < 0$ cases can be observed. Hence :

- for a rarefaction wave, there is a point where u is getting null and diverges out from this point
- for a compression wave, there is a point where u is getting very large because it is converging from its neighbour

Definition 2. *The point where A is changing sign is called a* **sonic point** *or a* **critical point**. *We prefer sonic point as "critical" can be used in other contexts.*

Remark 11. *There should be no confusion between A which is the speed of the wave and any variable u of the system which could be a fluid velocity. Nonetheless, for a neutral fluid in 1-dimension (the simpler case for the Euler equations), the three modes propagate at $u + c_s$, $u - c_s$ and u. For this last mode (also called the entropic mode) we have $A = u$ so a sonic point arise wherever $u = 0$.*

1.2.2 Dependence and influence domains

The dependence and influence domains are a fundamental property of hyperbolic equations. These domains are bounded by the caracteristic curves. The number of these caracteristics obviously depends on the number of equations, which is also the number of unknowns for a well-posed problem. This concept is illustrated below with 2 caracteristics as straight lines (with negative and positive slopes).

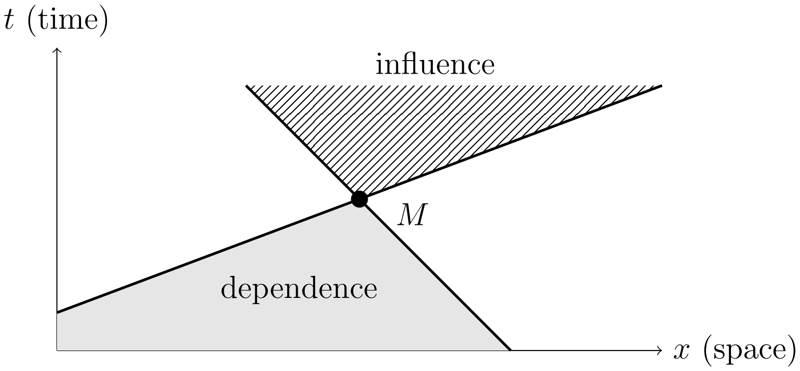

Figure 1.2: Schematics of the dependence (gray) and influence (hatched) domainsi at the M point.

In Fig. 1.2, the gray area depicts the dependence domain and the hatched region depicts the influence domain at the M point.

Definition 3. *The dependence domain associated to (x, t) is the region of phase space $\chi(\tau)$ so that $u(x, t)$ depends on all the $u(\chi, \tau)$.*

This concept is important because when space and time will be descritized, the choice of the scheme and the associated needed parameters will define a numerical domain of dependence. Obviously, numerical dependence domain and physical dependence domain are related.

Definition 4. *The influence domain associated to (x, t) is the region of phase space $\chi(\tau)$ so that $u(x, t)$ influences all the $u(\chi, \tau)$ values.*

1.2.3 Conservative and non-conservative forms

Let's consider the system of conservative equations for $u(x, t)$ of the form

$$\partial_t u + \partial_x f(u) = 0 \tag{1.9}$$

with $u(x, t = 0) = u_0(x)$. This form is called **consevative form** (or also divergence form).

Definition 5. *The non-linear fonction $f(u)$ is called the* **flux function**, *associated to $u(x, t)$. In the numerical context, we generally call it the* **physical flux function**, *to make a clear distinction with the* **numerical flux function**.

Notation 2. *In order to save tedious vector notations for the multidimensionnal case, we keep the scalar notation of the one-dimensionnal case. It means that for a n order system of conservative laws, u is a vector of length n, x is a vector of length 1, 2 or 3 and $f(u)$ is also a vector of length n (for a well-posed problem).*

Property 4. *At a sonic point, the flux function $f(u)$ has a null derivative (with respect to u).*

Assuming that the solution $u(x,t) \in C^1$ the u-derivative of the flux function can be calculated from the Jacobian matrix of f with respect to u. The resulting system has then the **non-conservative** form

$$\partial_t u + A(u)\partial_x u = 0 \tag{1.10}$$

still with $u(x, t = 0) = u_0(x)$. We then have $A = \partial_u f$.

For the multi-dimensional case, this equation writes in the vectorial form

$$\partial_t \mathbf{u} + \mathbf{A}(\mathbf{u}) \,.\, \partial_{\mathbf{r}} \mathbf{u} = 0 \tag{1.11}$$

The matrix $\mathbf{A}$ is the Jacobian of f, hence defined (in index notation) as

$$A_{ij} = \partial_{u_j} f_i(u) \tag{1.12}$$

It can be shown that for $u \in \mathbb{R}$, $\mathbf{A}$ is diagonalizable and all the eigen values are **real**. Hence, $\mathbf{A}$ writes

$$\mathbf{A} = \mathbf{Q} \,.\, \mathbf{\Lambda} \,.\, \mathbf{Q}^{-1} \tag{1.13}$$

where $\mathbf{Q}$ (the transition matrix) is a constant matrix of same dimension as $\mathbf{A}$ whose columns $\mathbf{r}_i$ are right caracteristic vectors of $\mathbf{A}$ and $\mathbf{Q}^{-1}$ is a constant matrix of same dimension as $\mathbf{A}$ whose rows $\mathbf{l}_i$ are left caracteristic vectors of $\mathbf{A}$. Then, $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements λ_i are the eigen values of $\mathbf{A}$.

The following change of variables

$$\mathbf{v} = \mathbf{Q}^{-1} \,.\, \mathbf{u} \tag{1.14}$$

defines the **caracteristic variables**. Then, Eq. (1.11) writes

$$\partial_t \mathbf{v} + \mathbf{\Lambda} \,.\, \partial_{\mathbf{r}} \mathbf{v} = 0 \tag{1.15}$$

1.3 One-dimensionnal Euler equations

1.3.1 Primitive variables and conservative variables

Notation 3. *For a given fluid, ρ is the mass density, $\rho\mathbf{v}$ is the momentum density and ρe[1] is the internal energy density*

The Euler equations are the equations describing an inviscid flow described by the three moments, ρ, $\rho\mathbf{v}$ and ρe. The term "moment" is justified as these 3 quantities are the moments[2] of order 0, 1 and 2 of the distribution functions. More precisely, the fluid quantities are the sum of each of these moments associated to each population constituting the plasma. The flow is inviscid as any viscous term would turn this system to be parabolic and its solution to be regularized by viscosity.

As usual, the hierarchy of fluid equations is a set of coupled equations where the n^{th} of this equation involves the $n+1$ order moment. Hence, to be well-posed, this system needs a "closure" equation : in the Euler equation, this is the equation of state for the kinetic pressure $p = p(\rho, e)$.

Definition 6. *This equation is a closure equation as the moment of order 2 is given by the lower order moments of order 0 and 1.*

Remark 12. *The moment of order 1 (the fluid velocity $\mathbf{v}$) is associated to the ram pressure, but has nothing to do with the kinetic pressure p. As a consequence, an EOS generally only depends on the density ρ.*

For the Euler equations, when deriving the fluid equations from the kinetic equation (Boltzmann, Vlasov...) we implicitely consider that the heat flux is null, that is the flow is adiabatic.

We introduce the Euler equations as they are the very classical equations usually considered when testing numerical schemes. But in a more general approach for plasma physics, the adiabatic hypothesis is not mandatory, and any kind of closure can be handled, provided the associated equation preserves the hyperbolicity of the system.

In the canonical case of a perfect monoatomic gas, the energy equipartition theorem gives $e = \frac{3}{2}nk_BT$ (with n being the particle density, that is $\rho = nm$ for particles of mass m), so that $p = \frac{2}{3}\rho e$.

Remark 13. *For the general case, any system has an adiabatic index γ (defined as the ratio c_p/c_v in thermodynamics). Hence, $p = (\gamma - 1)\rho e$.*

The Euler equations can be written in different forms, depending on the choice of the variables to describe the fluid (or plasma) : the **conservative variables** or the **primitive variables**.

Definition 7. *The conservative variables are the ones used in the conservative form of an hyperbolic system. For the Euler equations, they are ρ, $\rho\mathbf{v}$ and e.*

[1] e is then the internal energy per mass unit

[2] for the 2^{nd} order moment, ρe is a centered moment

Definition 8. *The primitive variables are the ones usually describing a physical system, that is ρ , $\mathbf{v}$ and p for a fluid.*

Remark 14. *The internal energy per mass unit e is the average value of the internal energy of the fluid in its frame. It means that the energy flow per mass unit does not contribute to e.*

Notation 4. *We note e_T the total energy density per mass unit, that is $e_T = e + \frac{1}{2}\mathbf{v}.\mathbf{v} = e + \frac{1}{2}v^2$ hence including the energy flow.*

1.3.2 Fluxes through surfaces

The equations governing the time evolution of a plasma are "local", meaning that the quantities that are followed through time are defined in a control volume. Like in classical transport theory, such a control volume is

- large enough so that the fluctuation level of any quantity is very small compared to the mean value of this quantity (that is a **mean value** has a **physical meaning**).
- small enough so that any macroscopic gradient at the scale of the control volume is very small compared to the **fluctuation level**.

The associated spatial scales greatly depends on the domain (from tenuous planetary magnetosphere to over-dense laboratory plasmas produced by high-power lasers) so will hence not be discussed here.

Notation 5. *We note V the control volume, $\mathrm{d}S$ the surface element of this control volume and $\mathbf{n}$ the associated unit vector normal to this surface and directed* **outward** *from this control volume.*

The change of mass density ρ, momentum density $\rho\mathbf{v}$ and internal energy density ρe is feeded by the flux of these quantities across $\mathrm{d}S$:

- the flux of mass density (for the mass transport) is $\rho\mathbf{v}.\mathbf{n}\mathrm{d}S$
- the flux of momentum density (for the momentum transport) is $(\rho\mathbf{v}\mathbf{v} + p\mathbf{1}).\mathbf{n}\mathrm{d}S$
- the flux of energy density (for the internal energy transport) is $(\rho e_T\mathbf{v} + p\mathbf{v}).\mathbf{n}\mathrm{d}S$

Notation 6. *The term $\mathbf{v}\mathbf{v}$ is the second order tensor reulting from the dyadic produc of $\mathbf{v}$ by itself and $\mathbf{1}$ is the second order unit tensor.*

As the reference problem, we then have for the conservative form of the one-dimensional Euler problem

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho v \\ \rho e_T \end{pmatrix} \qquad \mathbf{f}(\mathbf{v}) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ \rho e_T v + p v \end{pmatrix} \tag{1.16}$$

1.3.3 Euler equations : integral form

The temporal change between arbitrary times t_1 and t_2 of a conservative variable inside the control volume V equals the associated flux of this quantity through its surface $\mathbf{n}\,\mathrm{d}S$. The mass conservation equation is then

$$\int_V \rho(\mathbf{r},t_2)\,\mathrm{d}V - \int_V \rho(\mathbf{r},t_1)\,\mathrm{d}V = -\int_{t_1}^{t_2}\oint_{\partial V} \rho\mathbf{v}.\mathbf{n}\,\mathrm{d}S\mathrm{d}t \tag{1.17}$$

which rewrites

$$\int_V [\rho(\mathbf{r},t_2) - \rho(\mathbf{r},t_1)]\mathrm{d}V + \int_{t_1}^{t_2}\oint_{\partial V} \rho\mathbf{v}.\mathbf{n}\mathrm{d}S\mathrm{d}t = 0 \tag{1.18}$$

and is the integral form of the continuity equation (mass conservation).

The two other equations are straightforwardly obtained for the momentum density $\rho\mathbf{v}$ and the energy density ρe_T with their associated flux.

$$\int_V [\rho\mathbf{v}(\mathbf{r},t_2) - \rho\mathbf{v}(\mathbf{r},t_1)]\mathrm{d}V + \int_{t_1}^{t_2}\oint_{\partial V} [\rho\mathbf{v}\mathbf{v} + p\mathbf{1}).\mathbf{n}\mathrm{d}S\mathrm{d}t = 0 \tag{1.19}$$

$$\int_V [\rho e_T(\mathbf{r},t_2) - \rho e_T(\mathbf{r},t_1)]\mathrm{d}V + \int_{t_1}^{t_2}\oint_{\partial V} [\rho e_T\mathbf{v}\mathbf{v} + p\mathbf{v}).\mathbf{n}\mathrm{d}S\mathrm{d}t = 0 \tag{1.20}$$

1.3.4 Euler equations : differential form

Eq. (1.18), (1.19) and (1.20) can be transformed in differential form. As an illustration, Eq. (1.18) can be divided by $t_2 - t_1$

$$\int_V \frac{\rho(\mathbf{r},t_2) - \rho(\mathbf{r},t_1)}{t_2 - t_1}\mathrm{d}V + \frac{1}{t_2 - t_1}\int_{t_1}^{t_2}\oint_{\partial V} \rho\mathbf{v}.\mathbf{n}\mathrm{d}S\mathrm{d}t = 0 \tag{1.21}$$

In the limit $t_1 \to t_2$ and assuming that the time derivative $\partial_t\rho$ exist for all $\mathbf{r}$ in the whole domain, Eq. (1.21) writes

$$\int_V \partial_t\rho(\mathbf{r},t)\mathrm{d}V + \oint_{\partial V} \rho\mathbf{v}.\mathbf{n}\mathrm{d}S = 0 \tag{1.22}$$

Assuming that $\boldsymbol{\nabla}.(\rho\mathbf{v})$ is defined, the Gauss theorem for this equation gives

$$\int_V \partial_t\rho(\mathbf{r},t)\mathrm{d}V + \int_V \boldsymbol{\nabla}.\rho\mathbf{v}(\mathbf{r},t)\mathrm{d}V = 0 \tag{1.23}$$

This equation hold for whatever (even small) control volume. The integrand has hence to be zero so the differential form of the mass transport equation is

$$\partial_t\rho + \boldsymbol{\nabla}.\rho\mathbf{v} = 0 \tag{1.24}$$

Remark 15. *One needs to keep in mind that the differential form of this equation needs the hypothesis $\rho(\mathbf{r},t) \in C^1(x,t)$ and $\mathbf{v} \in (\mathbf{r},t)C^1(x,t)$. This is of importance in the cases of shocks or discontinuities.*

In the same way, the second Euler equation is

$$\partial_t \rho \mathbf{v} + \boldsymbol{\nabla}.(\rho \mathbf{v}\mathbf{v} + p\mathbf{1}) = 0 \tag{1.25}$$

and the third one is

$$\partial_t \rho e_T + \boldsymbol{\nabla}.(\rho e_T \mathbf{v} + p\mathbf{v}) = 0 \tag{1.26}$$

Remark 16. *In this derivation of the Euler equation, we assumed that the kinetic pressure of the plasma is isotrop, meaning that the pressure tensor reduces to $p\mathbf{1}$. In some cases (like in magnetized plasma), the media is no more isotropic, so that the diagonal terms of the pressure tensor are not all equal to the third of its trace. In the more general case of an agyrotropic plasma (in which the 2 directions perpendicular to the magnetic field are not equivalent), the pressure tensor can also have off-diagonal terms.*

1.4 Extensions of the Euler equations

The examples given below are not an extensive list of physical cases, but are just here in order to point-out how they can be treated, and what are the problems that could eventually arise.

1.4.1 Viscosity

Viscosity is generally introduced via transport theory, that is involving transport coefficient. The analytical form of this coefficient is out of scope of this introduction. In he second and third Euler equations, the term $-p\mathbf{1}$ is replaced by the stress tensor $\mathbf{T}$ defined as

$$\mathbf{T} = -p\mathbf{1} + \mu[\boldsymbol{\nabla}\mathbf{v} + \mathbf{v}\boldsymbol{\nabla} - \frac{2}{3}(\boldsymbol{\nabla}.\mathbf{v})\mathbf{1}] \tag{1.27}$$

In this equation, we hence need to introduce a dynamical viscosity μ. As a consequence, the divergence of these flux contains second order derivatives of the velocity, so that the 2^{nd} Euler equation turns to be parabolic.

1.4.2 Gravity

Gravity is generally introduced through the gravitational potential Φ which general definition is

$$\Phi(\mathbf{r}) = -G \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}' \tag{1.28}$$

In the case of the Earth (which mass is $M_\oplus$)

$$\Phi(\mathbf{r}) = -G\frac{M_\oplus}{r} \tag{1.29}$$

The second Euler equation then writes

$$\partial_t \rho\mathbf{v} + \boldsymbol{\nabla}.(\rho\mathbf{v}\mathbf{v} + p\mathbf{1}) = -\rho\boldsymbol{\nabla}\Phi \tag{1.30}$$

and the third one is unchanged, provided that the total energy now writes

$$e_T = e + \frac{1}{2}v^2 + \Phi \tag{1.31}$$

Hence, Eq. (1.30) has a "true" source term, that is a term that can not be explicitely written as a flux. This has important consequences on the choice of a scheme.

1.4.3 Magnetic field

When considering magnetic fields, the set of Euler equations is insufficient, as we also need an equation for the time evolution of the magnetic field. This one is of course the Maxwell-Faraday equation, which also involves the electric field. Such a system can then be complicated as the electric field can be associated to charge separation, meaning that we need to keep the fluid equations for each specie of the plasma.

The simpler case is the Magneto-Hydro-Dynamic one, where the plasma is described by a single fluid. With the quasi-neutral assumption, the electric field is lost from these equations, and then we need to keep the momentum equation of the electrons. For a collisionless plasma, neglecting the electron inertia, this equation reduces (in the MHD limit... that is very small values of both k and ω) to the ideal Ohm's law

$$\mathbf{E} = -\mathbf{v} \times \mathbf{B} \tag{1.32}$$

so that the Maxwell-Faraday equation reduces to

$$\partial_t \mathbf{B} + \boldsymbol{\nabla}.(\mathbf{v}\mathbf{B} - \mathbf{B}\mathbf{v}) = 0 \tag{1.33}$$

which pleasantly writes in conservative form.

In the flux of the momentum and energy equations, we then need to introduce the Maxwell stress tensor $\mathbf{T}_B$ defined as

$$\mathbf{T}_B = \frac{1}{\mu_0}[\mathbf{B}\mathbf{B} - \frac{B^2}{2}\mathbf{1}] \tag{1.34}$$

so the second Euler equations writes

$$\partial_t \rho\mathbf{v} + \boldsymbol{\nabla}.(\rho\mathbf{v}\mathbf{v} + p\mathbf{1} - \mathbf{T}_B) = 0 \tag{1.35}$$

The total energy density now contains a contribution from the magnetic field

$$e_T = e + \frac{1}{2}v^2 + \frac{1}{2}\frac{B^2}{\rho} \tag{1.36}$$

and the third Euler equation writes

$$\partial_t \rho e_T + \boldsymbol{\nabla}.(\rho e_T \mathbf{v} + p\mathbf{v} - \mathbf{T}_B.\mathbf{v}) = 0 \tag{1.37}$$

1.4.4 Radiations

As generally introduced in statistical physics, photons are a gas of bosons that can be described macroscopically by a set of thermodynamic quantities. We then introduce the radiative energy flux

$$\mathbf{F}_{\mathrm{rad}} = \int_{\mathbb{R}^+} \mathrm{d}\nu \oint_{\Omega} \mathrm{d}\omega \, \mathbf{n} I_\nu(\mathbf{r}, \mathbf{n}, \nu) \tag{1.38}$$

which depends on the spectral intensity I_ν. This term can (very roughly !) be considered as the "distribution function" of the photons.

The second thermodynamic quantity is the radiative pressure ; it is the pressure exerted on a surface by the massless photons as they are nonetheless carying a momentum.The radiative pressure is then

$$\mathbf{P}_{\mathrm{rad}} = \frac{1}{c}\int_{\mathbb{R}^+} \mathrm{d}\nu \oint_{\Omega} \mathrm{d}\omega \, \mathbf{n}\mathbf{n} I_\nu(\mathbf{r}, \mathbf{n}, \nu) \tag{1.39}$$

The second Euler equation is then

$$\partial_t \rho \mathbf{v} + \boldsymbol{\nabla}.(\rho \mathbf{v}\mathbf{v} + p\mathbf{1} + \mathbf{P}_{\mathrm{rad}}) = 0 \tag{1.40}$$

and the third one is

$$\partial_t \rho e_T + \boldsymbol{\nabla}.[(\rho e_T \mathbf{v} + p)\mathbf{v} + \mathbf{F}_{\mathrm{rad}}] = 0 \tag{1.41}$$

— 2 —

Features of numerical resolution

2.1 First attempts

2.1.1 Simple ordinary differential equation (ODE)

Consider the simple ODE for the function $y(t)$

$$\mathrm{d}_t y = -\alpha y \tag{2.1}$$

with $\alpha \in \mathbb{R}^+$ and the initial value $y(t=0) = y_0$. This linear first order ODE has the simple analytical solution

$$y(t) = y_0 \mathrm{e}^{-\alpha t} \tag{2.2}$$

Notation 7. *For time discretization, we use an exponents notation.*

Considering discret time steps $t^n = n\Delta t$, the most simple discretization is to approximate the time derivative at time t^n by the finite difference $\Delta t \mathrm{d}_t y(t^n) \sim y^{n+1} - y^n$, hence the scheme

$$y^{n+1} = y^n - \alpha y^n \Delta t \tag{2.3}$$

This **explicit Euler scheme** is the one obtained if y is approximated by a piecewise linear function. We can investigate how the solution behave depending on the Δt value.

Definition 9. *The solution of an ODE given by a finite difference method is stable if $|y^n| < \infty$ for $n \to \infty$.*

Definition 10. *The solution of an ODE given by a finite difference method is positive if $y^n > 0, \forall n$.*

We can easliy show that for the scheme given by Eq. (2.3)

- the criterion for **stability** is $\alpha \Delta t < 2$
- the criterion for **positivity** is $\alpha \Delta t < 1$

These values are neither universal (whatever the ODE) nor magical... it is just that a too large time-step badly resolves the curvature of the solution. The simple form of Eq. (2.3) leads to an explicit form of the ratio y^{n+1}/y^n, so the stability and positivity criterions depend on how the ratio y^{n+1}/y^n compares to 1 and 0, respecitively.

— Positive means that the condition $y^n > 0$ imply $y^{n+1} > 0$. As $y^{n+1} = y^n(1 - \alpha\Delta t)$, the scheme preserves positivity for $1 - \alpha\Delta t > 0$.
— Stable means that $y^{n+1}/y^n < 1$. For $1 - \alpha\Delta t > 0$, it is satisfied if $\alpha > 0$, which is part of the hypothesis. For $1 - \alpha\Delta t < 0$, it is satisfied for $\alpha\Delta t < 2$.

Remark 17. *Using this scheme, accurate results require a very small time-step value. This strong constraint can be avoided by using an higher order scheme, a scheme with adjustment of the time step, or a scheme specially adapted to this ODE.*

While apparently approximative, the foundation of this approach will be established in a more robust way in the next section when defining the order of a scheme.

2.1.2 Parabolic PDE

Let's now turn to the heat equation

$$\partial_t u = D\partial^2_{x^2} u \tag{2.4}$$

with $D \in \mathbb{R}^+$.

Remark 18. *The positivity of D is very important ; a negative diffusion coefficient would mean anti-diffusion. Hence, instead of the regularizing effect of the diffusion operator, anti-diffusion would increase in a dramatic way any gradients whatever its smallness.*

We introduce a spatial uniform discretization of grid-size Δx.

Notation 8. *For discrete variables, we use indices for the spatial discretization and exponent for the temporal discretization. Hence, $u^n_j = u(j\Delta x, n\Delta t)$.*

Remark 19. *For the sake of simplification, the origin of both space and time are zero.*

Whatever the initial and boundary conditions, the approximations $\Delta t\partial_t u(x_j, t^n) \sim u^{n+1}_j - u^n_j$ and $\Delta x^2\partial^2_{x^2} u(x_j, t^n) \sim u^n_{j+1} - 2u^n_j + u^n_{j-1}$, gives the explicit Euler scheme

$$u^{n+1}_j = u^n_j + \eta(u^n_{j+1} - 2u^n_j + u^n_{j-1}) \tag{2.5}$$

with the new (dimensionless) unknown $\eta = D\Delta t/\Delta x^2$. We have two criterions for this scheme :

- stability for $\eta < \frac{1}{2}$
- positivity for $\eta < \frac{1}{4}$

In the case of a PDE, whatever the number of values involved in the scheme, these criterions are not as straightforward as in the ODE case, because we have more than one unknown in the scheme for a single equation. Hence, the solution is to chose an analytical (local) approximation of the solution... piecewise, linear, exponential.

Such a diffusion equation has the clear effect of "smoothing out" all kinds of structures. To say it more roughly, a diffusion operator on a mathematical function acts as "sun on butter".

As a consequence, when the stability criterion is satisfied, all small-scale structures will be wiped out, the stiffest the gradients, the faster their smoothing. In the other cases, when the stability criteron is not satisfied, spurious oscillations appear and grow with time.

Remark 20. *The definition of η clearly shows that even if the Euler scheme is simple, the constraint on its stability is very expensive as the time step depends on the square of the grid size. Hence, a good spatial resolution will only be obtained at the expense of an expensive computation. Fortunately, some better schemes exist... but will not be detailed in this introductory course.*

2.1.3 Linear advection equation

In a pedagogical perspective, we discuss the most simple hyperbolic equation

$$\partial_t u + A\partial_x u = 0 \tag{2.6}$$

with $A = \text{const}$. The explicit **centered Euler scheme** is the one for which the spatial discretization writes

$$2\Delta x \partial_x u(x_j, t^n) \sim u^n_{j+1} - u^n_{j-1} \tag{2.7}$$

Notation 9. *We introduce the new parameter $\nu = \dfrac{\Delta t}{\Delta x}$.*

We can then write the centered Euler scheme

$$w^{n+1}_j = w^n_j - \frac{1}{2} A\nu [w^n_{j+1} - w^n_{j-1}] \tag{2.8}$$

Notation 10. *The numerical solution of a finite difference scheme is not a discretization of the exact solution u of the differential problem. We hence call w the exact solution of the discret problem (for a given scheme). That is $w^n_j \not\equiv u(x_j, t^n)$.*

It happens that whatever the smallness of the grid-size and time-step, and whatever the $A\nu$ value, this scheme is unconditionally unstable.

2.2 Taylor serie expansion and polynomial interpolation

2.2.1 Taylor series for derivative approximation

A numerical scheme can be obtained by following a quite simple procedure, usually called "finite difference method". To do so, one approximates the derivatives involved in the given PDE by finite difference by using a Taylor serie. As an illustration, for a function $u(x)$ discretized on $x_j = j\Delta x$,

$$u_{j+1} = u_j + \Delta x u'_j + \frac{\Delta x^2}{2} u''_j + O(\Delta x^3) \tag{2.9}$$

so we straightforwardly obtain

$$u'_j = \frac{u_{j+1} - u_j}{\Delta x} + O(\Delta x) \tag{2.10}$$

We deliberately put the remainders coming from the Taylos serie in the O function.

Definition 11. *The finite difference given by Eq. (2.11) is of order 1 because this is the power of Δx in the O function.*

Remark 21. *It is straightforward that the development at u_{j-1} around u_j would then give the alternative form*

$$u'_j = \frac{u_j - u_{j-1}}{\Delta x} + O(\Delta x) \tag{2.11}$$

From Eq. (2.9), we could also obtain a finite difference expression of order 2, by cancelling the u''_j term. For that purpose, we would write the same expression as Eq. (2.9) at u_{j-1}, and by differenciation, we would then obtain

$$u'_j = \frac{u_{j+1} - u_{j-1}}{2\Delta x} + O(\Delta x^2) \tag{2.12}$$

Remark 22. *While apparently promising, this finite difference is totally unstable and exhibit all kind of problems... Nonetheless, it deserves some attention for pedagogical reasons.*

The process is quite clear : Use a Taylor serie at an appropriate location on the grid and the combination in order to cancel-out low order terms. This approach also gives you the opportunity to "decide" which point is involved in the scheme. As an illustration, you can show that

$$u'_{j+1} = \frac{u_{j-1} - 4u_j + 3u_{j+1}}{2\Delta x} + O(\Delta x^2) \tag{2.13}$$

In order to get a second order scheme, we need to write 2 Taylor series :

$$u_{j-1} = u_{j+1} - 2\Delta x u'_{j+1} + \frac{(2\Delta x)^2}{2} u''_{j+1}$$

$$2u_j = 2u_{j+1} - 2\Delta x u'_{j+1} + \Delta x^2 u''_{j+1}$$

To cancel-out the Δx^2 term, we sum the first equation with 2 time the second. The result is straightforward.

2.2.2 Polynomial interpolation/extrapolation for reconstruction function

When we numerically solve a PDE, we work on sampled values that approximate at a given location (on the set of grid points) the exact solution. Using these sampled values, it can be interesting to have access to a functional representation.

Definition 12. *Any function created from samples is called a "reconstruction".*

Definition 13. *Any reconstruction passing through the sample points is called "interpolation" inside the domain of the samples.*

Definition 14. *Any reconstruction passing through the sample points is called "extrapolation" when it is used outside of the domain of the samples.*

There is a unique N^{th}-order polynomial passing through any set of $N+1$ samples. Such polynomial can be easy to reconstruct (like the Lagrange form or the Newton form), but not that easy to manipulate (to derive for example). They can also be more difficult to reconstruct (the Taylor series form) but then much simpler to use. While not giving details on how to reconstruct the coefficients a_i in a Taylor serie, we remind its form

$$p_N(x) = a_0 + a_1(x-b) + \cdots + a_N(x-b)^N \tag{2.14}$$

2.3 Scheme caracterizations and properties

2.3.1 Discret problem, truncation error and scheme order

We saw that Eq. (2.6) can be approximated by Eq. (2.8).

Notation 11. *We note $D(u) = 0$ the general differential problem and $\Xi(w) = 0$ the associated discret problem. Hence, D is a* **differential form***, Ξ a* **finite difference scheme***, u the excat solution of the differential problem and w the exact solution of the discrete problem.*

For the scheme given by Eq. (2.8), the Ξ operator acting on the $u(x,t)$ function is then

$$\Xi(u) = \frac{u_j^{n+1} - u_j^n}{\Delta t} + A\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \tag{2.15}$$

with $u_j^n = u(x_j, t^n) = u(j\Delta x, n\Delta t)$. All the terms in the right-hand-side of Eq. (2.15) can be written using a Taylor expansion around u_j^n. Using a Taylor-Lagrange formulation of the serie expansion, we have

$$u_j^{n+1} = u_j^n + \Delta t \partial_t u(x_j, t^n) + \frac{\Delta t^2}{2}\partial^2_{t^2} u(x_j, t^\star) \tag{2.16}$$

with $t^\star \in [t^n, t^{n+1}]$. This expansion is said to be at first order because the first order term is the last one that is explicited ($t^\star$ is not equivocally defined). In the same spirit, for the $u_{j\pm1}^n$ terms at second order,

$$u_{j\pm1}^n = u_j^n \pm \Delta x \partial_x u(x_j, t^n) + \frac{\Delta x^2}{2}\partial^2_{x^2} u(x_j, t^n) \pm \frac{\Delta x^3}{6}\partial^3_{x^3} u(x_\dagger, t^n) \tag{2.17}$$

with $x_\dagger \in [x_j, x_{j\pm1}]$. From Eq. (2.15), (2.16) and (2.17), we have

$$\begin{aligned}\Xi[u(x_j, t^n)] &= \partial_t u(x_j, t^n) + \frac{\Delta t}{2}\partial^2_{t^2} u(x_j, t^\star) + A\left[\partial_x u(x_j, t^n) + \frac{\Delta x^2}{3}\partial^3_{x^3} u(x_\dagger, t^n)\right] && (2.18)\\ &= [\partial_t u|_j^n + A\partial_x u|_j^n] + \frac{\Delta t}{2}\partial^2_{t^2} u(x_j, t^\star) + \frac{a\Delta x^2}{3}\partial^3_{x^3} u(x_\dagger, t^n) && (2.19)\end{aligned}$$

The first bracket in the right-hand-side of Eq. (2.19) is null as it satisfes the discrete problem $D(u) = 0$. This equation then writes

$$\Xi(u) = D(u) + T(x,t) \tag{2.20}$$

Definition 15. *In Eq. (2.20), $T(x,t)$ is the* **truncation error**.

Because we have $D(u) = 0$, the truncation error writes

$$T(x,t) = \frac{\Delta t}{2}\partial^2_{t^2} u(x_j, t^\star) + \frac{A\Delta x^2}{3}\partial^3_{x^3} u(x_\dagger, t^n) \tag{2.21}$$

If we now consider that $u(x,t)$ is C^2 for t and C^3 for x, then the derivative in Eq. (2.21) can be bounded. For that issue, we note M_{tt} the upper bound for $|\partial^2_{t^2} u(x,t)|$ and M_{xxx} the upper bound for $|\partial^3_{x^3} u(x,t)|$. Then, we have

$$T(x,t) \leq \frac{1}{2}M_{tt}\Delta t + \frac{A}{3}M_{xxx}\Delta x^2 \tag{2.22}$$

In the expression of $T(x,t)$, we can have Δt and Δx as small as we want, meaning that for a given problem and a given scheme, we can (in principle) make $T(x,t)$ as small as we want.

Definition 16. *Given a numerical scheme, if the associated truncation error writes $T(x,t) = A\Delta t^N + B\Delta x^M$ with $(A,B) \in \mathbb{R}^2$, the scheme is said to be of order N in time and M in space.*

Remark 23. *In Eq. (2.22), the terms that are explicited are the leading one because Δt and Δx being small, higher power of these terms are even smaller.*

2.3.2 Modified equation, numerical diffusion and dispersion

We solve the discrete problem $\Xi(u) = 0$. Because of Eq. (2.20), we then have

$$D(u) = -T(x,t) \tag{2.23}$$

If we only keep the leading term in the truncation error (the first one in Eq. (2.21)), u being a solution of Eq. (2.6), we have $\partial^2_{t^2} u = A^2 \partial^2_{x^2} u$ so the truncation error writes

$$T(x,t) = \frac{A^2 \Delta t}{2} \partial^2_{x^2} u(x,t) \tag{2.24}$$

As a result, the scheme $\Xi(w) = 0$ does not provide a solution of the initial differential problem $D(u) = 0$, but rather gives a solution of the modified equation $D(u) = -T(x,t)$.

Definition 17. *The modified equation of a numerical scheme is the differential equation that takes into account the (lowest order) terms of the truncation error.*

For the linear advection problem given by Eq. (2.6), the modified equation is then

$$\partial_t u + A \partial_x u = -\frac{A^2 \Delta t}{2} \partial^2_{x^2} u(x,t) \tag{2.25}$$

Remark 24. *The right-hand side of Eq. (2.25) is a diffusion term with a diffusion coefficient $-A^2 \frac{\Delta t}{2}$, meaning that as this term is negative, it gives rise to anti-diffusion. Any fluctuations are growing with time. This scheme is unstable as we already saw (without understanding why).*

Definition 18. *A scheme is said to be diffusive/anti-diffusive if the modified equation contains a diffusive/anti-diffusive term.*

For a better scheme, we generally still have a diffusion term, but with a positive coefficient. This coefficient is generally a function of Δt, Δx and A. While generally small, this is a diffusive term, so it acts on the numerical solution. It is then important to find a scheme that minimizes this term, or find a way to limit (or counter-balance) its growth.

Definition 19. *When the modified equation of a numerical scheme is of the form*

$$\partial_t w + A \partial_x w = \mu \partial^3_{x^3} w \tag{2.26}$$

the scheme is said to be dispersive.

The Fourier transform (both in space and time) of Eq. (2.26)

$$\widetilde{w}(k,\omega) = \int_{\mathbb{R}} w(x,t) \mathrm{e}^{\imath(kx - \omega t)} \mathrm{d}x \mathrm{d}t \tag{2.27}$$

has a solution satisfying $\omega = Ak + \mu k^3$. The physical interpretation of this equation is that the structures of w will propagate at a phase speed $\omega/k = A + \mu k^2$. The phase speed depending on k, all the Fourier modes will propagate at a speed depending on k. This will then give rise to dispersion, meaning that the form of a wave packet will not be preserved with time.

Remark 25. *When we established the modified equation, we gave the explicit form of the first term to emphasize what is dissipation and dispersion. In most cases, the modified equation of a given scheme contains both diffusion and dispersion.*

2.3.3 Stability, consistency, convergence

Definition 20. *We call Ω_j the set of j-indices so that $x_j \in \mathcal{D}$. We call $\mathcal{D}$ the definition domain (in space) of the function $u(x,t)$.*

There are three fundamental properties associated to the numerical resolution of a PDE.

Definition 21. *A numerical scheme is stable if $|w_j^n| < \infty$, $\forall j \in \Omega_j$ and $t \to \infty$.*

Definition 22. *A numerical scheme is consistent if $D \to \Xi$ for $\Delta x \to 0$ and $\Delta t \to 0$.*

Definition 23. *A numerical scheme converges if $|u_j^n - w_j^n| \to 0$, $\forall j \in \Omega_j$ and $t \to \infty$.*

The wide range of schemes (only some of them are discussed in this introduction) are all consistent. It is so because they were elaborated with the constraint that the associated truncation error are at least of order 1 in both space and time.

The stability is the minimal condition that one can expect from a numerical scheme ; Obviously, if its solution w_j^n is diverging, it is also diverging from the exact solution u of the differential problem $D(u) = 0$.

Definition 24. *The* **error** *of a scheme is defined at each grid point (and each time step) as $e_j^n = |u_j^n - w_j^n|$.*

A good scheme is hence a scheme that converges. A convergent scheme is then a scheme for which the error goes to zero, meaning that the discrete solution w_j^n is a good approximation of the exact solution $u(x_j, t^n)$.

The way we generally caracterize the error of a scheme will be given in the next subsection of this chapter. While the convergence of a scheme is not always easy to demonstrate, the Lax equivalence theorem is generally at the heart of demonstrating the convergence of a scheme.

Theoreme 1. *(Lax equivalence theorem). Given a well-posed initial value problem and a finite-difference approximation, that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.*

2.3.4 CFL condition

Because of the finite speed of waves, hyperbolic PDEs have a finite physical domain of dependence. For the simple advection problem given by Eq. (2.6) with the initial condition $u(x, t = 0) = u_0(x)$, the solution is $u(x, t) = u_0(x - At)$. This is so because $u(x, t)$ is constant along a caracteristic, which equation is in this case $d_t x = A$. As an example, such a caracteristic is depicted in Fig. 2.1 in solid line for a constant $A > 0$.

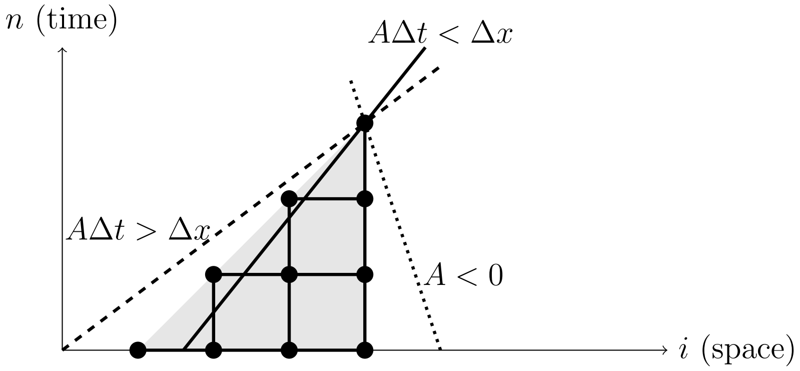

Figure 2.1: Example of caracteristics which lie or not in the numerical domain of dependance.

Consider also for the sake of illustration a non-centered scheme of the form

$$w_j^{n+1} = \gamma w_{j-1}^n + \beta w_j^n \tag{2.28}$$

for $(\gamma, \beta) \in \mathbb{R}^2$, meaning that w_j^{n+1} depends on w_{j-1}^n and w_j^n. In the same way, this set of points (w_{j-1}^n, w_j^n) depends on the three points w_{j-2}^{n-1} w_{j-1}^{n-1} and w_j^{n-1}. Going up to $n = 0$ we then define the numerical domain of dependance, represented in gray in Fig. 2.1.

Because the solution is in this case $u(x, t) = u_0(x - At)$, it is clear that the stability of a scheme cannot be guaranted if the caracteristic does not lie in the numerical domain of dependence. This is the CFL condition, from the paper [Courant et al., 1928]. In Fig. 2.1, the two caracteristics in dashed line (for a larger A value) and dotted line (for negative A value) do not satisfy the CFL condition :

- for the dashed line, Δx is too small or Δt too large, so that $a\Delta t > \Delta x$
- for the dotted line, A is negative. To satisfy the CFL condition, we should better have a scheme of the form $w_j^{n+1} = \gamma' w_{j+1}^n + \beta' w_j^n$

Property 5. *To satisfy the CFL condition, the full numerical domain of dependence must contain the physical domain of dependence.*

Remark 26. *The CFL condition is a necessary, but insufficient condition, for the stability of a scheme.*

Property 6. *The CFL condition is not $A\Delta t < \Delta x$ in the general case. For example, for a scheme with the five points at $j-2$, $j-1$, j, $j+1$ and $j+2$, the CFL condition is $A\Delta t < 2\Delta x$.*

Definition 25. *Using the notation already introduced $\nu = \dfrac{\Delta t}{\Delta x}$, the* **CFL number** *is $A\nu$.*

Because of the structure and properties of the advection equation, the value $u(x_j, t^{n+1})$ is inherited from its "first" ancestors, that is the ones defined at t^n. If all the ancestors are in a given range of values, then $u(x_j, t^{n+1})$ also lie in this range. This results in an important property that is to be satisfied by a scheme :

Property 7. *The* **upwind range condition** *states that*

$$\min_{x_{j-1}\le x\le x_j} u(x,t^n) \le u(x_j,t^{n+1}) \le \max_{x_{j-1}\le x\le x_j} u(x,t^n) \qquad \textit{for} \qquad 0 \le \eta A(x_j,t^{n+1}) \le 1 \tag{2.29}$$

$$\min_{x_j\le x\le x_{j+1}} u(x,t^n) \le u(x_j,t^{n+1}) \le \max_{x_j\le x\le x_{j+1}} u(x,t^n) \qquad \textit{for} \qquad -1 \le \eta A(x_j,t^{n+1}) \le 0 \tag{2.30}$$

$$\tag{2.31}$$

2.3.5 Stability analysis

If we ignore the effects of the boundary conditions, we can investigate the stability of a scheme by Fourier analysis. For that purpose, we use the Fourier modes

$$w_j^n = (\lambda)^n \mathrm{e}^{\imath(kj\Delta x)} \tag{2.32}$$

for a given k-mode. Of course, considering the boundary conditions would give some constraints on the discrete k-mode that can exist in the domain.

We did not demonstrate that this form is a solution. But this form is convenient, essentially because $w_j^{n+1} = \lambda w_j^n$, so a stability criterion is clearly $|\lambda| < 1$. If we consider the scheme given by Eq. (2.8), we straightforwardly obtain

$$\lambda(k) = 1 - \imath A\nu \sin k\Delta x \tag{2.33}$$

then, $|\lambda| > 1$, for all mesh ratio and almost all modes, meaning that this scheme, is always unstable : the discrete numerical solution w_j^n will unboundly grow for increasing time.

While we saw that the scheme given by Eq. (2.8) is definitively very bad for the linear advection equation given by Eq. (2.6), we will conclude this chapter with a new one, with much better performance while very simple : the upwind scheme.

The basic idea of this scheme is to use two points in order to discretize the spatial derivative. But in order to satisfy the CFL condition, these points depend on the sign of A. This scheme then writes

$$w_j^{n+1} = \begin{cases} w_j^n - A\nu(w_{j+1}^n - w_j^n) & \text{for } A < 0 \\ w_j^n - A\nu(w_j^n - w_{j-1}^n) & \text{for } A > 0 \end{cases} \tag{2.34}$$

Remark 27. *This scheme was proposed in a 1952 paper [Courant et al., 1952] and can also be called the CIR method from the names of its authors.*

The upwind scheme can be written in a single line using the quantities $\frac{1}{2}(|A| - A)$ and $\frac{1}{2}(|A| + A)$. The name of this scheme is straightforward, as it uses the points in the direction from which the wind (that is the wave speed) is blowing.

On can then easily make the stability analysis of the upwind scheme and then obtain for the amplification coefficient

$$\lambda(k) = 1 - 2A\nu(1 - A\nu)(1 - \cos k\Delta x) \tag{2.35}$$

which gives

$$|\lambda|^2 = 1 - 4A\nu(1 - A\nu)\sin^2 \frac{1}{2}k\Delta x \tag{2.36}$$

It then follows that $|\lambda(k)| \leq 1$ for all k, provided that the CFL condition $|A\nu| < 1$ is satisfied. The upwind scheme is then stable when the CFL condition is satisfied.

Remark 28. *While the upwind scheme is of order 1, that is smaller than the order of the explicit centered Euler scheme, it is a stable scheme. As a consequence, the order of a scheme is not a guarantee of its stability.*

In Eq. (2.32), space and time indices are not treated in the same way. We could have used a Fourier serie in time as we did in space. But the form that is used with λ is enough for stability analysis. Obviously, $\lambda \in \mathbb{C}$, so that its modulus quantify how instable is the scheme, and its phase quantify its dispersion. More precisely, for a perfect one-dimensional plane wave, we should have $\lambda = \mathrm{e}^{\imath\omega n\Delta t}$, then giving its phase for an ω-mode.

2.4 Spatial and temporal discretization

2.4.1 Grid deposition

Finite difference. With the finite difference approach, each quantity involved in the equations to be solved (that is the components of u and the components of the associated flux f) are **discretized** on a structured grid, generally uniform. With such an approach, we then have

$$w_j^n \sim u(j\Delta x, n\Delta t) \tag{2.37}$$

that is, using words : w_j^n is a discret set of values on a structured uniform grid, which approximates the exact solution (which is a continuous function) sampled on this grid.

Definition 26. *The scheme is said to be implicit if w_j^{n+1} depends at least from another w value also defined at time step $n+1$.*

Finite volume. For this grid deposition, we also consider a spatial discretization on $N+1$ nodes uniformally distributed with a mesh $\Delta x = L/N$. In the same way, the time is discretized with a time-step Δt. For the finite volume approach, the conservative laws are **integrated** on a elementary mesh $[x_{j-1/2}, x_{j+1/2}]$. Then, the integration of Eq. (1.9) gives

$$\partial_t \int_{x_{j-1/2}}^{x_{j+1/2}} u \, \mathrm{d}x + \int_{x_{j-1/2}}^{x_{j+1/2}} \partial_x f \, \mathrm{d}x = 0 \tag{2.38}$$

As a consequence, such approach is dealing with the average value $\overline{u}_j^n$ of the continuous function $u(x,t)$ calculated on the mesh $[x_{j-1/2}, x_{j+1/2}]$ defined as

$$\overline{u}_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, n\Delta t) \, \mathrm{d}x \tag{2.39}$$

With such a grid deposition, w_j^n is hence an approximation of $\overline{u}_j^n$. Following the structure of Eq. (2.38) and the spatial derivative in the second integral, this equation simply writes

$$\partial_t w_j^n + \frac{F_{j+1/2}^n - F_{j-1/2}^n}{\Delta x} = 0 \tag{2.40}$$

The physical flux has also to be approximated :

- because in Eq. (2.40), u is approximated by w, and its time derivative is approximated by finite difference
- because un Eq. (2.40), the flux has to be calculated at the cell interface where u (that is w) is not defined. hence, some interpolation/extrapolation is needed, meaning that the numerical flux is not equal to the physical one

Hence, F_j^n is the **numerical flux** which approximates the **physical flux** $f[u(j\Delta x, n\Delta t)]$. The power of this approach is that the flux are then defined on the interface of a cell which is an important property of the associated conservative law.

Property 8. *In the finite volume approach, two adjacent cells share the same flux : the outgoing flux of a given cell is then exactly the one ingoing in the next adjacent cell.*

Remark 29. *It is then clear that a finite volume scheme is consistent if $\lim_{\Delta t \to 0\ ,\ \Delta x \to 0} F_j^n = f(j\Delta x, n\Delta t)$.*

The power of the finite volume approach is also that, because of the integration in a cell, the spatial derivative of the physical flux has disappeared.

Definition 27. *We call* **weak formulation** *the way to rewrite a PDE in a way that no derivative of the solution appears.*

While surprising, PDE can have solution which derivative is not necessarily defined. The finite volume formalism is hence well-suited to handle the weak solution of a PDE when discontinuities are initially imposed or are rising with time integration.

Definition 28. *We call* **weak solution** *of a PDE the function that satisfies the integral form of the original PDE.*

Remark 30. *The derivative of the weak solution of a PDE is not necessarily defined !*

One should remember that in the first chapter, the derivative form of the Euler equations was obtained from the integral form. The integral form was "simply" a balance in a small (but finite) control volume. The derivative form was obtained in the limit of a control volume size going to zero. While it mathematically makes sense, it is physically a nonsense because as already discussed, the control volume has to be large enough so that one can define a macroscopic average value.

Remark 31. *Some schemes are obtained with a grid deposition of finite-difference type. Nevertheless, some of them can also be obtained with a different grid deposition, namely using finite volume. It is then important to keep in mind that in these cases (see for example the classical linear Lax-Wendroff scheme), the discret values w_j^n are not defined in the same way.*

Whatever finite difference or finite volume approach, w_j^n is an approximate value of the continuous function $u(x, t)$ at $x_j = j\Delta x$ and $t^n = n\Delta t$.

2.4.2 Time integration

The temporal derivative is generally obtained using an explicite Euler scheme of the form

$$\partial_t w_j^n = \frac{w_j^{n+1} - w_j^n}{\Delta t} \tag{2.41}$$

The time step Δt has to be defined in order to satisfy the CFL condition (for an explicit scheme). For that issue, we generally need to calculate the spectral radius of the Jacobian matrix $\partial_u f$, that is the largest of the absolute values of its eigen values. As a consequence, in codes where the time step is initially defined, it is important to be sure that this condition will be fulfilled during its time integration.

The conservative form of any finite volume scheme is then generally of the form

$$w_j^{n+1} = w_j^n + \frac{\Delta t}{\Delta x}[F_{j+1/2}^n - F_{j-1/2}] \tag{2.42}$$

meaning that when the form of the numerical flux is the only one provided, it has to be used in Eq. (2.42). In the other cases, the expression of w_j^{n+1} is provided, as well as all the numerical flux needed.

2.4.3 Stencils diagrams

A finite volume scheme generally gives the form of the numerical flux $F^n_{j+1/2}$ as a function (linear or not) of the quantities w^n_j, w^n_{j+1}.... Hence, in Eq. (2.42), it means that the value of w^{n+1}_j depends on a given set of values of w^n. This set can then be graphically represented in a stencil diagram as the one given in Fig. 2.2

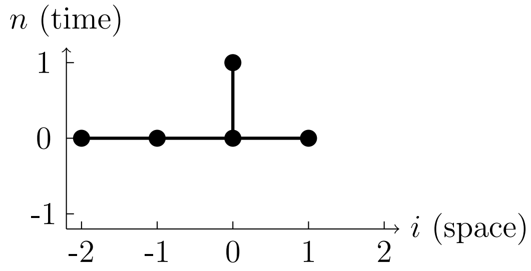

Figure 2.2: Example of stencil diagram.

In this example, w^{n+1}_j depends on w^n_{j-2}, w^n_{j-1}, w^n_j and w^n_{j+1}. The **width** of this (asymmetric) stencil is 4.

Definition 29. *The* **width of a stencil** *is the number of points involved at time t^n to calculate w^{n+1}_j.*

There is a terminology associated to these stencils that simplifies the naming of several classes. For the spatial discretization,

- **Forward Space** means that the derivative at grid point j uses a set of w_{j+k} points with $k \in \mathbb{N}^+$
- **Backward Space** means that the derivative at grid point j uses a set of w_{j-k} poionts with $k \in \mathbb{N}^+$
- **Centered Space** means that the derivative at grid point j uses a difference on indices symetric around j (for example $j+1$ and $j-1$). A centered scheme then uses an odd number of points.

The same terminology is used for the time discretization.

Example 4. *The finite difference scheme given by Eq. (2.8) is FTCS (Forward Time Centered Space).*

Example 5. *The finite difference scheme given by Eq. (2.34) is FTBS.*

A last word on stencils : it appears that in some schemes, the stencil is not constant, but depends on the numerical solution w^n_j.

Definition 30. *An* **adaptative sencil** *is a stencil wich width and structure depends on the local values of the numerical solution.*

2.4.4 Ghost cells

The width of a stencil is important for the appropriate treatment of the boundary conditions. In the example of Fig. 2.2, for any points on the left border of the domain, the scheme needs two points that will be out of the domain. In the same way, on the right border, the scheme will need one point out of the domain.

While these points are needed, but out of the domain, it means that we then need some extra cells, out of the domain, in order to have these discret values of w.

Definition 31. *The cell points out of the domain but needed to apply a stencil are called* **ghost cells**.

For periodic simulations, the values of thes w_j^ns in the ghost cells are simply a copy of the appropriate values of the w_j^ns that are inside the domain. For non-periodic simulations, we then need to set some dedicated boundary condition, and then "translate" these physical conditions to deduce the values of the w_j^n in the ghosts cells. The most classical boundary conditions for a plasma are **perfectly conducting wall** and **dielectric wall**, where one can deduce the values of the electric field, magnetic field, charge density...

- In a perfectly conducting wall, the charges are free to move, so that $\mathbf{E} = 0$.
- In a dielectric wall, there are no free charges, so that $\mathbf{\nabla}.\mathbf{E} = 0$, that is $\mathrm{d}_n E_n = 0$. The normal component of $\mathbf{B}$ is also continuous (as it is divergence-free), that is $\mathbf{\nabla} \times \mathbf{E} = 0$ for the 2 tangential components, which is satisfied for $\mathbf{E}_T = 0$.

Remark 32. *The treatment of boundary conditions can be very touchy... meaning that in many cases, this task is more an art than a science !*

2.4.5 Artificial viscosity

As already saw, the first derivative (at second order) can be approximated as

$$\partial_x f|_{x_j} = \frac{1}{2\Delta x}[f(x_{i+1}) - f(x_{j-1})] + O(\Delta x^2) \tag{2.43}$$

Such an approximation gives awful results in a finite-difference scheme. The reason is that in Eq. (2.43), the derivative at odd indices j will only depends on fluxes calculated at even indices (and vice versa). This lead to a separation between the odd- and even-indices. Such a phenomenon is usually called "odd-even $2\Delta x$-wave oscillations".

We saw in Eq. (2.25) that the FTCS scheme exhibits an anti-diffusive term that make it unstable. This outlines that this kind of finite difference is associated to a (negative) diffusion term. As an alternative, Eq. (2.13) gives the derivative at x_{j+1}.

$$\partial_x f|_{x_{j-1}} = \frac{1}{2\Delta x}[-f(x_{i+1}) + 4f(x_j) - 3f(x_{j-1})] + O(\Delta x^2) \tag{2.44}$$

that we call the "second order forward difference". This derivative can be written in a different way as

$$\partial_x f|_{x_{j-1}} = \frac{f(x_j) - f(x_{j-1})}{\Delta x} - \frac{\Delta x}{2} \frac{f(x_{i+1}) - 2f(x_j) + f(x_{j-1})}{\Delta x^2} + O(\Delta x^2) \tag{2.45}$$

In Eq. (2.45), the first term approximates a first order derivative (forward difference) and the second term approximates a second order derivative (centered difference). Such a second order derivative appears in the Navier-Stokes equation, and while the difference with the Euler equation is a viscous term, the second term in Eq. (2.45) is called artificial viscosity.

Remark 33. *It has to be clear that the artificial viscosisity generally arises from the way a scheme is build, and has absolutely nothing to do with the physical viscosity.*

Definition 32. *Any second, fourth, sixth... and other even-order differences in a modified equation are called* **artificial viscosity***.*

Definition 33. *Any third, fifth... and other odd-order differences in a modified equation are called* **artificial dispersion***.*

From a general point of view, a viscous-like term in an advection equation add a term on the right hand side of the conservation law as

$$\partial_t u + \partial_x f = \partial_x[\epsilon(u)\partial_x u] \tag{2.46}$$

with $\epsilon \geq 0$. This last condition is essential. With $\epsilon \leq 0$, such a term is associated to **anti-diffusion**, and drives fluctutions growth at an unacceptable level deadly fast.

A conservative FTCS discretization of Eq. (2.46) writes

$$\frac{w_j^{n+1} - w_j^n}{\Delta t} + \frac{1}{2\Delta x}[f(w_{j+1}^n) - f(w_{j-1}^n)] = \frac{1}{\Delta x}\left[\epsilon_{j+1/2}^n \frac{w_{j+1}^n - w_j^n}{\Delta x} - \epsilon_{j-1/2}^n \frac{w_j^n - w_{j-1}^n}{\Delta x}\right] \tag{2.47}$$

Rearanging these terms, the numerical flux (which is used in the conservative form of the advection equation) can be written as the sum of two terms

$$F_{j+1/2}^n = \frac{1}{2}[f(w_{j+1}^n) + f(w_j^n)] - \frac{\epsilon_{j+1/2}^n}{2}(w_{j+1}^n - w_j^n) \tag{2.48}$$

- The first term is widely encountered in finite volume method ; the numerical flux at $x_{j+1/2}$ is simply an average of the physical flux at x_j and x_{j+1}.
- The second term clearly contains a first order derivative of w. While in Eq. (2.40) the numerical flux appears in a first order derivative, this second term is clearly a viscous term associated to a diffusion operator.

The artificial viscosity then appears as a flux correction.

Remark 34. *Eq. (2.45) can be written as Eq. (2.48) with $\epsilon_{j+1/2} = -\Delta x/2$.*

While the viscosity ϵ_j^n is strictly numerical, some schemes can use this formalism. Hence, the game is to find the proper way to define the value of the artificial viscosity. This one has to be

- as small as possible in smooth region where no smoothing is needed
- just large enough in shocks region in order to prevent a too large steepening of the wave front that could turn the scheme unstable

Then, ϵ_j^n should be a function of differences involving w_j^n's values. This will be discussed later on in the techniques of flux limiters.

2.5 The Riemann problem

The Riemann problem (for the Euler equations) is the one associated to Eq. (2.6) with the initial condition

$$u(x, t = 0) = u_0(x) = \begin{cases} u_L & \text{for } x < 0 \\ u_R & \text{for } x > 0 \end{cases} \tag{2.49}$$

The Riemann problem has an exact analytical solution for the Euler equations, for any scalar conservation laws, as well as for any linear system of equations. Furthermore, the system is *self-similar* (or *self-preserving*).

Definition 34. *A PDE depending on x and t coordinates is* **self-similar** *if its solution only depends on the ratio x/t rather than on x and t separately.*

A consequence is that the solution is constant along any lines $x = \eta t$ (with $\eta \in \mathbb{R}$) passing through the origin in the (x, t) plane. The Riemann problem is hence the simplest test-case for numerical approximation of discontinuities.

Property 9. *By self-similarity, numerical methods using Riemann solvers require only the flux along $x = 0$.*

2.5.1 Weak solution

Even in the cases where the flux function $f(u)$ as well as the initial conditions $u_0(x)$ are C^1, discontinuities can arise meaning that the solution $u(x, t)$ of the problem could eventually not be C^1. Such **weak solutions** are physically describing shocks or discontinuities.

The differential form of the conservation equation is not mathematically adapted, to treat shocks and discontinuities. We then have to choose the integral form, meaning that these equations are integrated over small control volume in space. Such an approach is closely related to the finite volume formulation of the conservative problem already introduced in section 2.4.1

Theoreme 2. *(Lax-Wendroff theorem). If a conservative numerical scheme for a hyperbolic system of conservation laws converges, then it converges towards a* **weak solution**.

2.5.2 The jump conditions

Consider Eq. (1.9) written in differential form. Then, consider a weak solution $u(x,t)$ that is discontinuous along a (regular) curve Γ in the (x,t) plan. Calling Ω_L and Ω_R the two domains separated by Γ where u_L and u_R (the restriction of u to Ω_L and Ω_R, respectively) are regular solutions of the problem. Let n be the normal to Γ, oriented from Ω_R to Ω_L. The jump conditions between $u(x,t)$ and its associated flux function $f(u)$ across the Γ curve writes

$$(u_R - u_L)n_t + [f(u_R) - f(u_L)]n_x = 0 \tag{2.50}$$

Eq. (1.9) writes $\mathbf{\nabla}.\mathbf{t} = 0$ with $\mathbf{\nabla} = (\partial_t, \partial_x)$ and $\mathbf{t} = (u, f)$. Then, Eq. $\mathbf{\nabla}.\mathbf{t} = 0$ can be integrated in the (x,t) plane, and using the Green-Stokes theorem, one gets $\mathbf{t}.\mathbf{n}|_L = \mathbf{t}.\mathbf{n}|_R$, that is $u_L n_t + f_L n_x = u_R n_t + f_R n_x$.

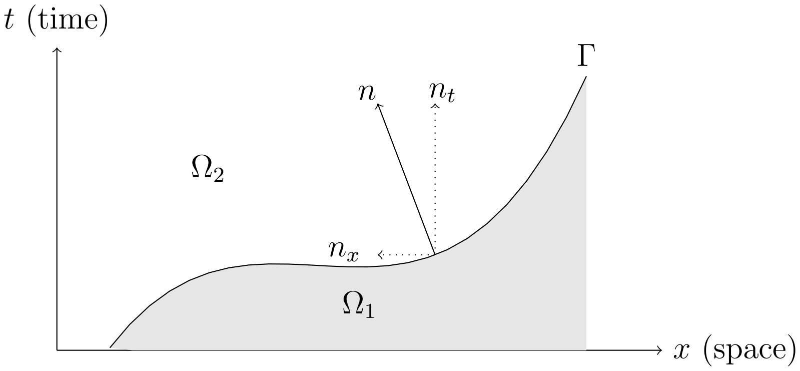

Figure 2.3: Weak solution across a discontinuity Γ.

Such a formulation is not very nice because the way to define the normal vectors (in order to obtain its two coordinates) is not that easy and the physical dimension of its coordinates are not the same. A way to solve this problem is to write Eq. (1.9) in integral form,

$$\int_{x_1}^{x_2} [u(x,t_2) - u(x,t_1)]\mathrm{d}x + \int_{t_1}^{t_2} [f(x_2,t) - f(x_1,t)]\mathrm{d}t = 0 \tag{2.51}$$

Considering a schock traveling at speed s, we can then choose x_1, x_2, t_1 and t_2 in order to follow the shock (as illustrated in Fig. 2.4), that is satisfying

$$x_2 - x_1 = s(t_2 - t_1) \tag{2.52}$$

We introduce the quantities $\Delta t = t_2 - t_1$ and $\Delta x = x_2 - x_1$. We also considere that for small enough Δt and Δx, $\int_{x_1}^{x_2} u(x, t_2)\mathrm{d}x = u_L \Delta x$ and $\int_{t_1}^{t_2} f(x_2, t)\mathrm{d}t = f_R \Delta t$. One can do the same at t_1 and x_1.

Notation 12. *For the sake of simplification, we note $f_q = f(u_q)$ where q can stand for a spatial index or a side index in the context of a discontinuity/shock.*

We hence obtain

$$\Delta x(u_L - u_R) + \Delta t(f_R - f_L) = 0 \tag{2.53}$$

As we noted $s = \Delta x/\Delta t$ the shock speed, we then obtain the Rankine-Hugoniot jump condition

$$s(u_R - u_L) = f_R - f_L \tag{2.54}$$

which is depicted in Fig. 2.4.

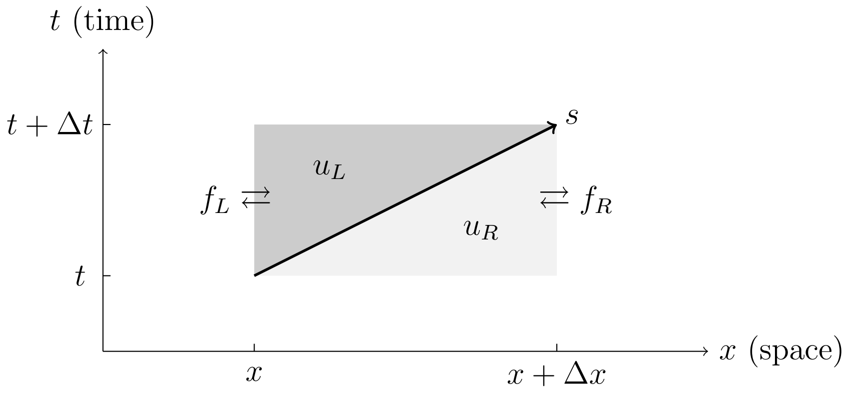

Figure 2.4: Jump across a discontinuity/shock.

This equation gives the shock speed from the conservative form as

$$s = \frac{f_R - f_L}{u_R - u_L} \tag{2.55}$$

2.5.3 Exact solution

Remember the Riemann problem in vectorial form

$$\partial_t \mathbf{u} + \mathbf{A} \,.\, \partial_x \mathbf{u} = 0 \tag{2.56}$$

$$\mathbf{u}(x, t=0) = \mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L & \text{for } x < 0 \\ \mathbf{u}_R & \text{for } x > 0 \end{cases} \tag{2.57}$$

for a constant $N \times N$ matrix $\mathbf{A}$. As we introduced $\mathbf{v}$ in Eq. (1.14), the advection equation for $\mathbf{v}$ is the same as the one for $\mathbf{u}$, provided that $\mathbf{A}$ is replaced by $\boldsymbol{\Lambda}$, $\mathbf{u}_L$ by $\mathbf{v}_L = \mathbf{Q}^{-1} \,.\, \mathbf{u}_L$ and $\mathbf{u}_R$ by $\mathbf{v}_R = \mathbf{Q}^{-1} \,.\, \mathbf{u}_R$. In other words, the problem for $\mathbf{v}$ is still a Riemann problem, but with a diagonal Jacobian.

The matrix $\boldsymbol{\Lambda}$ is diagonal with three eigen values λ_i in the case $N = 3$ (which is the case for the one-dimensional Euler equations). As an example, we report the 3 caracteristics $\mathrm{d}_t x = \lambda_i$ in Fig. 2.5.

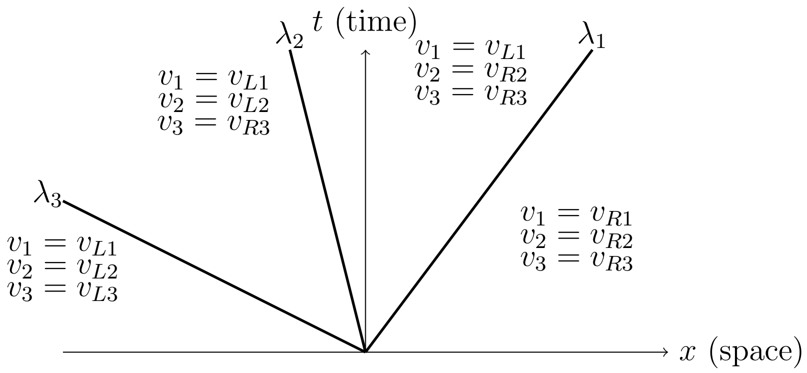

Figure 2.5: Solution of the Riemann problem for a linear system of 3 partial differential equations.

Let $\Delta v_i = v_{Ri} - v_{Li}$ be the jump in the i^{th} variable defined as $\Delta \mathbf{v} = \mathbf{v}_R - \mathbf{v}_L = \mathbf{Q}^{-1} \,.\, \Delta \mathbf{u}$.

We introduce a specific notation : $\Delta \mathbf{v}_i$ is the column vector which i^{th} component is Δv_i, all the other components being null. Then,

$$\Delta \mathbf{u} = \mathbf{Q} \,.\, \Delta \mathbf{v} = \mathbf{r}_i \Delta v_i \tag{2.58}$$

where $\mathbf{r}_i$ is the i^{th} column of $\mathbf{Q}$. Each vectors $\mathbf{r}_i$ are then the right eigen vector associated to $\mathbf{A}$. We then get the full form of $\mathbf{u}(x/t)$, that is the solution of the Riemann problem.

— 3 —

Overview of classical schemes

Many of the schemes presented in this chapter lie on the conservative form of the advection given by Eq. (2.42) :

$$w_j^{n+1} = w_j^n + \nu[F_{j+1/2}^n - F_{j-1/2}^n] \tag{3.1}$$

with $\nu = \Delta t/\Delta x$.

3.1 Lax-Wendroff type schemes

The simplest approach to set the value of the numerical flux $F_{j+1/2}^n$ could be the FTFS scheme with the flux

$$F_{j+1/2}^n = f(w_{j+1}^n) \tag{3.2}$$

With such a flux, small-scale oscillations grow even faster than for the naive scheme (the Euler explicit scheme given by Eq. (2.8)) then making this scheme useless.

A first alternative would be to choose a centered form for the numerical flux like

$$F_{j+1/2}^n = \frac{1}{2}[f(w_{j+1}^n) + f(w_j^n)] \tag{3.3}$$

which is also inoperant : even its fully implicit form (hence with a BTCS stencil) couple the two disadvantages of smearing out heavily any large scale structures but also let somme wiggles appear.

Remark 35. *The main idea of all the Lax-Wendroff type schemes is to provide a correction of the flux given by Eq. (3.3) using the artificial viscosity form already suggested in Eq. (2.48).*

Notation 13. *These types of schemes can be called "first generation" schemes.*

3.1.1 Lax-Wendroff scheme

This scheme by [Lax and Wendroff, 1960] is based on a Taylor expansion in **time**, up to the third order. As a result, the Lax-Wendroff scheme is $O(\Delta x^2, \Delta t^2)$. In the linear case of a physical flux defined as $f(x,t) = Au(x,t)$ where A is a real constant, the numerical flux writes

$$F_{j+1/2}^n = \frac{1}{2}[f(w_{j+1}^n) + f(w_j^n)] - \frac{A^2\nu}{2}[w_{j+1}^n - w_j^n] \tag{3.4}$$

associated to the stencil displayed in Fig. 3.1.

Remark 36. *In the non-linear case, we need to find an appropriate form for $A = \partial_u f$. Different forms have been proposed, all constituting the wide class of Lax-Wendroff type scheme.*

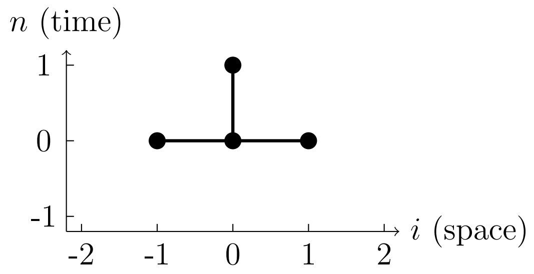

Figure 3.1: Stencil diagram of the Lax-Wendroff scheme.

For initial conditions having strong gradients or even discontinuities, this scheme produces overshoots rising very quickly, even if they generally does not much grow with time. Such a scheme is hence usefull for smooth initial conditions, providing that no stiff gradients appear later.

This Lax-Wendroff scheme can be generalized to the nonlinear cases. It is then necessary to replace the A value in Eq. (3.4) by a numerical approximation of a non-local $A^n_{j+1/2}$ value.

3.1.2 Lax-Friedrichs scheme

The flux of the Lax-Friedrichs scheme by [Lax, 1954] writes

$$F^n_{j+1/2} = \frac{1}{2}[f(w^n_{j+1}) + f(w^n_j)] - \frac{1}{2\nu}[w^n_{j+1} - w^n_j] \tag{3.5}$$

hence, reporting this expression in Eq. (3.3) gives the stencil displayed in Fig. 3.2.

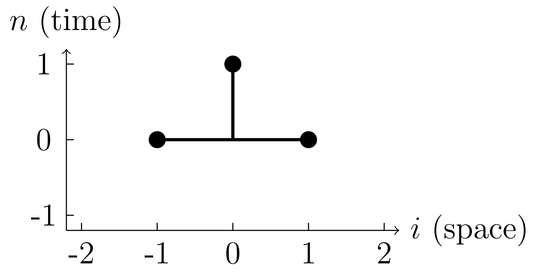

Figure 3.2: Stencil diagram of the Lax-Friedrichs scheme.

The lax-Friedrichs scheme has the same kind of disadvantage of the FTFS scheme and is hence never used. But [Richtmyer, 1962] proposed a 2-steps form using a (staggered) predictor-corrector for w^n_j :

$$w^{\star}_{j+1/2} = \frac{w^n_{j+1} + w^n_j}{2} - \frac{1}{2}\nu[f(w^n_{j+1}) - f(w^n_j)] \tag{3.6}$$
$$w^{n+1}_j = w^n_j - \nu[f(w^{\star}_{j+1/2}) - f(w^{\star}_{j-1/2})] \tag{3.7}$$

Later, [macCormack, 1969] improved this scheme using a FS and BS centering in a non-staggered predictor-corrector to obtain :

$$w^{\star}_j = w^n_j - \nu[f(w^n_{j+1}) - f(w^n_j)] \tag{3.8}$$
$$w^{n+1}_j = \frac{w^n_j + w^{\star}_j}{2} - \nu[f(w^{\star}_j) - f(w^{\star}_{j-1})] \tag{3.9}$$

3.1.3 Beam-Warming scheme

The Beam-Warming scheme by [Beam and Warming, 1976] is also $O(\Delta x^2, \Delta t^2)$ with the flux

$$F^n_{j+1/2} = \frac{1}{2}[3f(w^n_j) - f(w^n_{j-1})] - \frac{A^2\nu}{2}[w^n_j - w^n_{j-1}] \tag{3.10}$$

associated to the stencil displayed in Fig. 3.3.

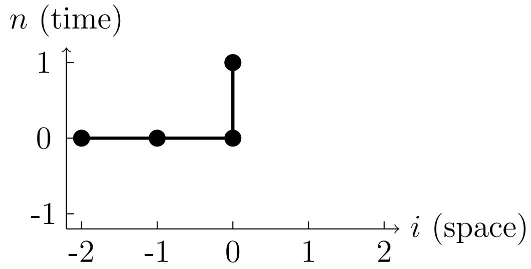

Figure 3.3: Stencil diagram of the Beam-Warming scheme.

In the non-linear case, it is necessary to find a local evaluation of A. It is clear that this scheme is decentered in the upwind direction. It suffer from the same overshoot structures as the Lax-Wendroff scheme, but interestingly, these overshoots appear for the Beam-Warming scheme in the opposite direction. This suggests the Fromm scheme.

3.1.4 Fromm scheme

The Fromm scheme by [Fromm, 1968] is then an average of the Lax-Wendroff and Beam-Warming scheme : hence, it is also of order $O(\Delta x^2, \Delta t^2)$ with a flux given by

$$F^n_{j+1/2} = \frac{1}{2}[F^{\mathrm{Lax-Wendroff}}_{j+1/2} + F^{\mathrm{Beam-Warming}}_{j+1/2}] \tag{3.11}$$

The results is very good for wave propagation, and also quite smooth for stiff gradients... so far the best scheme. The stencil diagramm is displayed in Fig. 3.4.

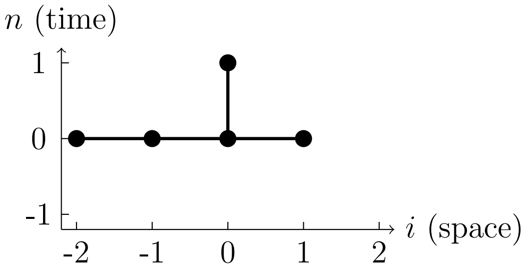

Figure 3.4: Stencil diagram of the Fromm scheme.

3.2 Riemann solvers & Flux Difference Splitting schemes

3.2.1 Approximate Riemann solvers

The solution of the Riemann problem needs to be treated in the vectorial case ; most of the subtilities are hidden in the scalar case.

The exact solution of the Riemann problem is a 4 steps procedure :

- calculate the diagonal matrix $\mathbf{\Lambda}$ from $\mathbf{A}$, the eigen values λ_i and the transition matrix $\mathbf{Q}$
- calculate the components of $\mathbf{v}_L = \mathbf{Q}^{-1} \,.\, \mathbf{u}_L$ and $\mathbf{v}_R = \mathbf{Q}^{-1} \,.\, \mathbf{u}_R$
- compare x/t to the λ_i's and deduce the v_i (hence the $\Delta\mathbf{v}_i$'s) using Fig. 2.5
- then get the solution $\Delta\mathbf{u}$ using Eq. (2.58)

Notation 14. *We note* $\mathbf{u}_0 \equiv \mathbf{u}(x = 0, t) = u(x/t = 0)$.

Property 10. *The solution* $\mathbf{u}(x, t)$ *of the Riemann problem is self-similar, meaning that it only depends on the ratio* x/t.

Hence, $\mathbf{u}$ is constant along any lines passing through $(x = 0, t = 0)$. The solution of $\mathbf{u}(x, t > 0)$ is then "connected" to $\mathbf{u}_0$ along the $x = 0$ line. Solving the Riemann problem is then restricted to studying the evolution of $\mathbf{u}$ along $x = 0$. As a consequence, the positive λ_i's will play a different role as the negative ones.

It is clear that the exact solution of the Riemann problem is very expensive as these solutions request to calculate $\mathbf{\Lambda}$, $\mathbf{Q}^{-1}$ and some of their products. Then, it, can hardly be achieved in a numerical code at the edges of each cells, and at each time steps. For that reason, many approximate Riemann solvers have been proposed and few of these are reported below.

Notation 15.

$$\lambda_i^- = \min(0, \lambda_i) \qquad\qquad \lambda_i^+ = \max(0, \lambda_i) \tag{3.12}$$

Note 1. *The main challenge in solving a conservative equation is to calculate a numerical flux* $\mathbf{F}_{i+1/2}$ *as a numerical approximation of the physical flux. As the numerical flux is calculated at the interface between 2 adjacent cells, we need to perform some interpolation. Remembering that* $\mathbf{f}(\mathbf{u}_2) \sim \mathbf{f}(\mathbf{u}_1) + \partial_{\mathbf{u}}\mathbf{f}.(\mathbf{u}_2 - \mathbf{u}_1) \equiv \mathbf{f}(\mathbf{u}_1) + \mathbf{A}.(\mathbf{u}_2 - \mathbf{u}_1)$*, the previous remark makes clear that most of the work lies in finding a good approximation of* $\mathbf{A}.\mathbf{u}_0$*.*

We can verify that along $x = 0$,

$$\mathbf{A} \, . \, \mathbf{u}_0 = \mathbf{A} \, . \, \mathbf{u}_L + \sum_{i=1}^{3} \mathbf{r}_i \lambda_i^- \Delta v_i = \mathbf{A} \, . \, \mathbf{u}_R - \sum_{i=1}^{3} \mathbf{r}_i \lambda_i^+ \Delta v_i \tag{3.13}$$

so by averaging, we have

$$\mathbf{A} \, . \, \mathbf{u}_0 = \frac{1}{2} \mathbf{A} \, . \, (\mathbf{u}_R + \mathbf{u}_L) - \frac{1}{2} \sum_{i=1}^{3} \mathbf{r}_i |\lambda_i| \Delta v_i \tag{3.14}$$

We define $\mathbf{\Lambda}^+$ and $\mathbf{\Lambda}^-$ the diagonal matrix with diagonal elements λ_i^+ and λ_i^-, respectively and $|\mathbf{\Lambda}|$ the diagonal matrix with diagonal elements $|\lambda_i|$. We then have

$$\mathbf{\Lambda} = \mathbf{\Lambda}^+ + \mathbf{\Lambda}^- \qquad\qquad |\mathbf{\Lambda}| = \mathbf{\Lambda}^+ - \mathbf{\Lambda}^- \tag{3.15}$$

We straightforwardly define

$$|\mathbf{A}| = \mathbf{Q} \, . \, |\mathbf{\Lambda}| \, . \, \mathbf{Q}^{-1} \tag{3.16}$$

and Eq. (3.14) takes the form

$$\mathbf{A} \, . \, \mathbf{u}_0 = \frac{1}{2} \mathbf{A} \, . \, (\mathbf{u}_R + \mathbf{u}_L) - \frac{1}{2} |\mathbf{A}| \, . \, (\mathbf{u}_R - \mathbf{u}_L) \tag{3.17}$$

The first term is quite straightforward because $\mathbf{A}.\mathbf{u}_R = \mathbf{F}_R$ and $\mathbf{A}.\mathbf{u}_L = \mathbf{F}_L$. But the second one is less obvious because $\mathbf{A}$ being non-linear, that is local, one needs to define "where" to calculate $|\mathbf{A}|$. The forthcoming subsections present a (reduced) set of methods to approximate $\mathbf{A}$ in Eq. (3.17).

3.2.2 Godunov schemes

The interested reader could check the very nice and brief review of the Godunov-type methods by [Sweby, 1999].

In the special case of scalar linear advection equation, each cell-average values (separated by the discontinuities) are "simply" advected (Lagrangian stage). But in an Eulerian perspective, the cell averaged advection equation is needed.

Let's consider the one-dimensional scalar conservation law

$$\partial_t u + \partial_x f = 0 \tag{3.18}$$

that we integrate over the domain "D" indicated by the shadow area in Fig. 3.5 for any given value $\alpha \in \mathbb{R}$ (that is positive or negative). The initial condition is still $n(x < 0, t = 0) = u_L$ and $u(x > 0, t = 0) = u_R$.

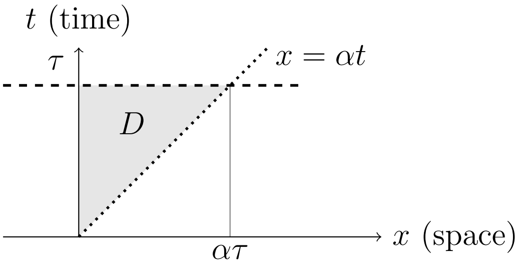

Figure 3.5: Illustration of the integral domain to determine the flux function.

This integration writes

$$\int_{x=0}^{\alpha\tau}\int_{t'=x/\alpha}^{\tau} \partial_t u \,\mathrm{d}t'\mathrm{d}x + \int_{t'=0}^{\tau}\int_{x=0}^{\alpha\tau} \partial_x f \,\mathrm{d}t'\mathrm{d}x = 0 \tag{3.19}$$

that is

$$\int_{x=0}^{\alpha\tau} [u(x,\tau) - u(x, x/\alpha)]\,\mathrm{d}x + \int_{t'=0}^{\tau} [f(u(\alpha t', t')) - f(u(0,t'))]\,\mathrm{d}t' = 0 \tag{3.20}$$

The solution of a Riemann problem being self-preserving, it is constant along $d_t x = \text{const.}$

$$u(0,t) = u_0 = \text{const.} \qquad\qquad u(x, x/\alpha) = u_1 = \text{const.} \tag{3.21}$$

so that Eq. (3.20) writes

$$\int_{x=0}^{\alpha\tau} [u(x,\tau) - u_1]\,\mathrm{d}x + \int_{t'=0}^{\tau} [f(u_1) - f(u_0)]\,\mathrm{d}t' = 0 \tag{3.22}$$

that is

$$f(u_0) - f(u_1) = \frac{1}{\tau}\int_{x=0}^{\alpha\tau} [u(x,\tau) - u_1]\,\mathrm{d}x \tag{3.23}$$

Scalar conservation laws preserve the monotonicity, meaning that if $u_R > u_L$ ($u_R < u_L$) at $t = 0$, then $u_R > u_L$ ($u_R < u_L$) $\forall t > 0$. We then have two cases :

- if $\forall x \leq \alpha\tau$ we have the order relation $u_1 \geq u(x,\tau) \geq u_0$, then $f(u_0) - f(u_1) \leq 0$, that is

$$f(u_0) \leq f(u_1) \tag{3.24}$$

Since $u_L \leq u_1 \leq u_R$ is arbitrary, u_1 can be any u value in the range $[u_L, u_R]$, so

$$f(u_0) = \min_{u_L \leq u \leq u_R} f(u) \tag{3.25}$$

- if $\forall x \leq \alpha\tau$ we have the order relation $u_1 \leq u(x,\tau) \leq u_0$, then $f(u_0) - f(u_1) \geq 0$, that is

$$f(u_0) = \max_{u_L \geq u \geq u_R} f(u) \tag{3.26}$$

Then, u_0 being the value of u at the origin by Eq. (3.21), the flux function at the origin writes

$$f(u(0,t)) = \begin{cases} \min\limits_{u_L \leq u \leq u_R} f(u) & \text{if } u_L < u_R \\ \max\limits_{u_L \geq u \geq u_R} f(u) & \text{if } u_L > u_R \end{cases} \tag{3.27}$$

Applying this formula for the flux function in the cell between j and $j+1$, one obtains the **Godunov's first-order upwind flux** by [Godunov, 1959]

$$\overline{F}^{\text{Godunov}}_{j+1/2}(w_j, w_{j+1}) = \begin{cases} \min\limits_{w_j^n \leq u \leq w_{j+1}^n} f(u) & \text{if } w_j < w_{j+1} \\ \max\limits_{w_j^n \geq u \geq w_{j+1}^n} f(u) & \text{if } w_j > w_{j+1} \end{cases} \tag{3.28}$$

Notation 16. *The overline on the numerical flux F is intended to outline that such a flux can be read as an "average" using the two points given as parameters.*

It is then clear that the trick is to calculate the min or max of the physical flux function in a given cell. This min/max is either at an endpoint of the interval for a monotonic physical flux function, or where the derivative of the flux function is null (sonic point). The Godunov numerical flux function is then in a simpler form

$$F^n_{j+1/2} = \begin{cases} \min[f(w_j^n), f(w_{j+1}^n), f(u^\star)] & \text{if } w_j < w_{j+1} \\ \max[f(w_j^n), f(w_{j+1}^n), f(u^\star)] & \text{if } w_j > w_{j+1} \end{cases} \tag{3.29}$$

where $u^\star$ refers to any and all sonic points between w_j and w_{j+1}.

Remark 37. *With any analytical form of the flux, the derivative $\partial_u f$ can be calculated, and for a gentle enough form of this flux, the $u^\star$ value at which this derivative is null can be analyticaly obtained.*

Example 6. *For the Burger equation, the physical flux writes $f(u) = \frac{1}{2}u^2$, so $u^\star = 0 = f(u^\star)$. Obviously, there is a $u^\star$ in a range $[w_j, w_{j+1}]$ if and only if $w_j.w_{j+1} < 0$.*

It si obvious that in the linear case of a physical flux given by $f(u) = au$, the Godunov first-order upwind scheme gives the classical FTBS upwind scheme. Of course, some more complicated form of this flux can be used in the class of the Godunov schemes.

Definition 35. *The* **entropy condition** *for a numerical scheme is equivalent to the second law of thermodynamics. This condition ensures that the entropy of a system numerically integrated through time is necessarily increasing or stay constant.*

The Godunov first-order upwind scheme satisfies the entropy condition of the Euler's equations and preserves the positivity of the variables. But this scheme costs a lot as it needs to solve the Riemann problem at each interface and for each time steps.

Theoreme 3. *(Godunov's theorem). Linear numerical schemes for solving partial differential equations (PDEs), having the property of not generating new extrema (monotone scheme), can be at most first-order accurate.*

This is an important theorem : for the general cases where some shocks or discontinuities can appear or live throught the time evolution of the system, any scheme with an order of accuracy larger than one will fail close to these shocks or discontinuities. On the other hand, a first-order accuracy scheme will have bad performances in smooth regions where complex coupling of mode can be at play. This is for this reason that most of efforts have later been dedicated to find non-linear schemes as they are the only ones ensuring both monotonicity preserving property and high order. Nonetheless, a proper way to manage the consequences of the Godunov theorem can also be to use a smart symbiosis of a first-order scheme with a larger order one. For that purpose, "smart" means that we should find a way to weight the associated fluxes in a sort of average.

3.2.3 Roe-first order upwind method

The main idea of this scheme by [Roe, 1981] is to find a linear approximation of the Jacobian matrix $\mathbf{A}$ in the quasi-linear form of the advection equation given by Eq. (2.56). For a scalar conservative equation, the Taylor expansions of the flux function around $\mathbf{u}_R$ and $\mathbf{u}_L$ are

$$\mathbf{f}(\mathbf{u}) \sim \mathbf{A}(\mathbf{u}_L)\,.\,(\mathbf{u}-\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_L) \qquad \text{or} \qquad \mathbf{f}(\mathbf{u}) \sim \mathbf{A}(\mathbf{u}_R)\,.\,(\mathbf{u}-\mathbf{u}_R) + \mathbf{f}(\mathbf{u}_R) \tag{3.30}$$

We hence would like to define a $\mathbf{A}_{RL}$ matrix such that

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = \mathbf{A}_{RL}\,.\,(\mathbf{u}_R - \mathbf{u}_L) \tag{3.31}$$

The solutions satisfying these N equations is not unique for the $N \times N$ matrix $\mathbf{A}_{RL}$ (with its N^2 free parameters). Roe proposed the simplified form

$$\mathbf{A}_{RL} = \mathbf{A}(\mathbf{u}_{RL}) \tag{3.32}$$

where the unknown $\mathbf{u}_{RL}$ lies between $\mathbf{u}_R$ and $\mathbf{u}_L$. The matrix $\mathbf{A}$ being known, solving Eq. (3.32) gives the $\mathbf{u}_{RL}$ components. For the one-dimensional Euler equations, the $\mathbf{A}_{RL}$ matrix is called the **Roe-average Jacobian matrix**.

Once the Roe-average conservative variables of the problem $\mathbf{u}_{RL}$ have been defined with Eq. (3.32), the Roe-average wave speed λ_i can be computed as well as the wave speed Δv_i. Finally, the computation of the conservative variables $\mathbf{u}$ needs also to compute the transition matrix $\mathbf{Q}$. Using the same notation as in Eq. (2.58) where $\mathbf{r}_i$ is the i^{th} column of $\mathbf{Q}$, the numerical flux of this method then writes

$$\mathbf{F}^n_{j+1/2} = \frac{1}{2}[\mathbf{f}(w^n_{j+1}) + \mathbf{f}(w^n_j)] - \frac{1}{2}\sum_{k=1}^{N} \mathbf{r}_k|\lambda_k|\Delta v_k \tag{3.33}$$

Remark 38. *Roe's approximate Riemann solver is somehow two and a half time less expensive than the exact Riemann solver.*

3.2.4 Upwind schemes

This scheme has already been introduced in Eq. (2.34). We will write it in a different way, just to illustrate the idea of flux vector splitting. In a pedagogical perspective, we treat the linear scalar advection equation given by Eq. (2.6). In the case $A \in \mathbb{R}^+$, the numerical flux for a BS centering is simply

$$F^n_{j+1/2} = f(w^n_j) = Aw^n_j \tag{3.34}$$

and we already saw that it is of order $O(\Delta t, \Delta x)$, consistent and stable with the CFL condition $A\Delta t \leq \Delta x$. For $A \in \mathbb{R}^-$, such an upwind scheme which is also called **donor cell** has the numerical flux

$$F^n_{j+1/2} = f(w^n_{j+1}) = Aw^n_{j+1} \tag{3.35}$$

For a more general case of a velocity A which can have both signs during the time evolution of the flow, the upwind scheme should then write

$$\frac{w^{n+1}_j - w^n_j}{\Delta t} + \left(\frac{A+|A|}{2}\right)\frac{w^n_j - w^n_{j-1}}{\Delta x} + \left(\frac{A-|A|}{2}\right)\frac{w^n_{j+1} - w^n_j}{\Delta x} = 0 \tag{3.36}$$

meaning that with the two new unknowns

$$A^+ \equiv \frac{A+|A|}{2} \qquad\qquad A^- \equiv \frac{A-|A|}{2} \tag{3.37}$$

we can define two fluxes

$$F^+_{j+1/2} = A^+ w^n_j \qquad\qquad F^-_{j+1/2} = A^- w_{j+1} \tag{3.38}$$

and the scheme writes

$$\frac{w_j^{n+1} - w_j^n}{\Delta t} + \frac{F_{j+1/2}^+ - F_{j-1/2}^+}{\Delta x} + \frac{F_{j+1/2}^- - F_{j-1/2}^-}{\Delta x} = 0 \tag{3.39}$$

which means that we could use the conservative form given by Eq. (3.1) with the numerical flux defined as

$$F_{j+1/2}^n = F_{j+1/2}^+ + F_{j+1/2}^- = A^+ w_j + A^- w_{j+1} \tag{3.40}$$

Remark 39. *The concept of "Flux vector splitting" now clearly appears ; the idea is to split the numerical flux in two parts, F^+ associated to positive flux, that is $\partial_u F^+ \geq 0$ and F^- associated to negative flux, that is $\partial_u F^- \leq 0$*

Finally, in artificial viscosity form, this flux writes

$$F_{j+1/2}^n = \frac{1}{2}[f(w_j^n) + f(w_{j+1}^n)] - \frac{|A|}{2}(w_{j+1}^n - w_j^n) \tag{3.41}$$

This form can only be applied to scalar equations. In the more general case, the game is to find, depending on the kind of waves, the most important(s) one(s) and retain the absoluite value of their velocity for $|A|$.

Notation 17. *The form of the flux given by Eq. (3.41) is clearly the same as the one given by Eq. (3.33) for a scalar equation, and will then be later called the Roe approximate Riemann solver.*

The two forthcoming subsection deal with the cases where A should be replaced by a Jacobian matrix,

3.2.5 Rusanov scheme

One remember that the Godunov flux has a general form qiven by Eq. (3.28). The form we choose to write this flux as $F_{j+1/2}^n = \overline{F}^{\text{Godunov}}(w_j^n, w_{j+1}^n)$ is quite clear : the solution of the Riemann problem a at the cell interface $i + \frac{1}{2}$ is $w_{i+1/2}^\star(x/t) = \text{RP}[w_j^n, w_{j+1}^n]$, so that the flux at this interface is $F_{j+1/2}^n = f[w_{j+1/2}^\star(0)] = F^\star(w_j^n, w_{j+1}^n)$.

The Rusanov scheme is then a way to write this averaged numerical flux as

$$F^\star(w_j^n, w_{j+1}^n) = \frac{1}{2}[f(w_j^n) + f(w_{j+1}^n)] - \frac{1}{2}A(w_j^n, w_{j+1}^n)(w_{j+1}^n - w_j^n) \tag{3.42}$$

with de definition of $A(w_j^n, w_{j+1}^n)$ as

$$A(w_j^n, w_{j+1}^n) = \max(|f'(w_j^n)|, |f'(w_{j+1}^n)|) \tag{3.43}$$

3.2.6 HLL scheme

This scheme is part of the familly of the two-waves schemes. Thes two waves have a wave speed given by $S_L = \min_{w_L,w_R} \min_i \lambda_i(w)$ and $S_R = \max_{w_L,w_R} \max_i \lambda_i(w)$. These two terms can also write

$$S_L = \min(w_L, w_R) - \max(A_L, A_R) \qquad S_R = \max(w_L, w_R) + \max(A_L, A_R) \tag{3.44}$$

Hence, the numerical flux is

$$\begin{aligned}
F^\star(w_j^n, w_{j+1}^n) &= f(w_j^n) \text{ for } S_L > 0 && (3.45)\\
F^\star(w_j^n, w_{j+1}^n) &= \frac{S_R f(w_j^n) - S_L f(w_{j+1}^n) + S_L S_R (w_j^n w_{j+1}^n - w_j^n)}{S_R - S_L} \text{ for } S_L < 0 \text{ and } s_R > 0 && (3.46)\\
F^\star(w_j^n, w_{j+1}^n) &= f(w_{j+1}^n) \text{ for } S_R < 0 && (3.47)\\
& && (3.48)
\end{aligned}$$

This form of the flux can be proven to result from the integral form of the equation with the two caracteristics S_L and S_R.

3.3 Flux limiter schemes

3.3.1 The big picture

In section 3.1.4, we saw that the Fromm scheme, defined as an average of the Lax-Wendroff and the Beam-Warming schemes had pretty good performances. Doing so, we tried to get the best from each scheme. But the average between these two fluxes is not case-sensitive, and might not always be the best linear combination we could hope.

The idea of flux limiter methods is to use a weighted average of two fluxes, one dedicated to correctly treat regular regions and another one dedicated to shocks and/or discontinuities. The associated weights in this average for a new time step should then adjust, depending on the gradients of the numerical solution obtained at the current time step.

In the scalar conservation law

$$w_j^{n+1} = w_j^n - \nu[F_{j+1/2}^n - F_{j-1/2}^n] \tag{3.49}$$

the flux could be defined as

$$F_{j+1/2}^n = F_{j+1/2}^{(1)} + \phi_{j+1/2}^n (F_{j+1/2}^{(2)} - F_{j+1/2}^{(1)}) \tag{3.50}$$

where $F_{j+1/2}^{(1)}$ and $F_{j+1/2}^{(2)}$ are the conservative numerical fluxes of two different methods.

Definition 36. *In Eq. (3.68), the parameter $\phi_{j+1/2}^n$ controlling the weight of the linear combination is called a* **flux limiter***.*

Remark 40. *In Eq. (3.68), the numerical flux is then given by the first term $F^{(1)}_{j+1/2}$, but then "limited" by the second term.*

In order to save computation, the spatial index of the flux limiter is oftenly bump, up or down, by one half. Hence, for $a > 0$, Eq. (3.68) writes

$$F^n_{j+1/2} = F^{(1)}_{j+1/2} + \phi^n_j (F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2}) \tag{3.51}$$

and for $a < 0$, Eq. (3.68) writes

$$F^n_{j+1/2} = F^{(1)}_{j+1/2} + \phi^n_{j+1} (F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2}) \tag{3.52}$$

The trick is of course to find an analytical expression for the ϕ function, as well as the appropriate unknown on which it applies, in order to find a solution-sensitive average. This being said, ϕ should be a function depending on the numerical solution w^n_j and more precisely on its derivative, approximated by finite differences.

As a general idea, we need to evaluate how close we are from a shock or from a wave. This means that we should certainly calculate a 1-grid cell difference, but also evaluate how it change from one cell to the directly adjacent one. To do so, we can introduce r^+_j and r^-_j, defined as

$$r^+_j = \frac{w^n_j - w^n_{j-1}}{w^n_{j+1} - w^n_j} \qquad r^-_j = \frac{w^n_{j+1} - w^n_j}{w^n_j - w^n_{j-1}} \tag{3.53}$$

Remark 41. *It is clear that $r^+_j = 1/r^-_j$.*

An alternative approach is to consider a shock indicator based on the numerical flux difference $F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2}$. In smooth regions, any of these numerical fluxes should have very comparable schemes, so this difference should be "small". But in shock regions, the order of accuracy of a scheme generally drops, meaning that such a difference should then be "large". Hence, the r parameters defined below could also be defined as

$$r^+_j = \frac{F^{(2)}_{j-1/2} - F^{(1)}_{j-1/2}}{F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2}} \qquad r^-_j = \frac{F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2}}{F^{(2)}_{j-1/2} - F^{(1)}_{j-1/2}} \tag{3.54}$$

Remembering the way numerical fluxes writes in term of artificial viscosity,

$$F^{(1)}_{j+1/2} = \frac{1}{2}[f(w^n_{j+1}) + f(w^n_j)] - \frac{1}{2}\epsilon^{(1)}_{j+1/2}(w^n_{j+1} - w^n_j) \tag{3.55}$$

$$F^{(2)}_{j+1/2} = \frac{1}{2}[f(w^n_{j+1}) + f(w^n_j)] - \frac{1}{2}\epsilon^{(2)}_{j+1/2}(w^n_{j+1} - w^n_j) \tag{3.56}$$

$$\tag{3.57}$$

so that the flux difference writes

$$F^{(2)}_{j+1/2} - F^{(1)}_{j+1/2} = \frac{1}{2}[\epsilon^{(2)}_{j+1/2} - \epsilon^{(1)}_{j+1/2}](w^n_{j+1} - w^n_j) \tag{3.58}$$

With such expression, we can write the ratios of flux difference, and study their properties. It appears that for $\epsilon^{(1)}_{j+1/2} > \epsilon^{(2)}_{j+1/2}$, $r^{\pm}_j \geq 0$ if the w^n_j's are monotonically increasing or decreasing, and $r^{\pm}_j \leq 0$ if the w^n_j's have a local maximum or a minimum. xxxiHence, there is genrerally a link between the sign of $r^{\pm}_j$ and the existence of a sonic point.

In the next subsections, we investigate various form of the $\phi(r)$ function and will try to emphasize the constraints on this function.

3.3.2 Van Leer flux limiter

The flux-limited method by [van Leer, 1974] for the linear advection equation with $a > 0$ is

$$w^{n+1}_j = w^n_j - \nu[F^n_{j+1/2} - F^n_{j-1/2}] \tag{3.59}$$

with the flux

$$F^n_{j+1/2} = \frac{1+\eta^n_j}{2} F^{\text{Lax}-\text{Wendroff}}_{j+1/2} + \frac{1-\eta^n_j}{2} F^{\text{Beam}-\text{Warming}}_{j+1/2} \tag{3.60}$$

with the new unknown η^n_j defined as

$$\eta^n_j = \frac{|r^+_j| - 1}{|r^+_j| + 1} \tag{3.61}$$

and

$$r^+_j = \frac{w^n_j - w^n_{j-1}}{w^n_{j+1} - w^n_j} \tag{3.62}$$

This formulation is intended to make the formulation clearer, while it differs from the one introduced in the previous subsection. At the end of this section, the form of various flux limiter will be given and compared in a graphical way.

3.3.3 Sweby flux limiter

We previously saw that the Roe-first order upwind scheme (see Eq. (3.41) for the scalar problem) and the Lax-Wendroff scheme (see Eq. (3.4) for the scalar problem) have complementary properties : the first one is doing well near jump discontinuities, whereas the last one does well in smooth region.

For the Sweby flux-limited scheme by [Sweby, 1984],

$$F^n_{j+1/2} = F^{\text{Roe}}_{j+1/2} + \phi^n_j [F^{\text{Lax}-\text{Wendroff}}_{j+1/2} - F^{\text{Roe}}_{j+1/2}] \tag{3.63}$$

There exist conditions on the Sweby's flux-limited function, as the one that satisfies the proper "upwinding" of the scheme, depending on the sign of A in Eq. (2.6).

Notation 18. *In the following developments, ϕ^n_j has to be read as the function $\phi(r^+_j)$ on the unknown r^n_j given by Eq. (3.62)*

There exist several flux limiters among which,

- **minmod** : $\phi(r) = \max[0, \min(1, r)]$
- **Chakravarthy & Osher** : $\phi(r) = \max[0, \min(\beta, r)]$ with $1 \leq \beta \leq 2$
- **Van Leer** : $\phi(r) = \frac{r+|r|}{1+r}$
- **Van Albada** : $\phi(r) = \max\left[0, \frac{r+r^2}{1+r^2}\right]$
- **SuperBee** : $\phi(r) = \max[0, \min(1, 2r), \min(2, r)]$

These flux limiters are displayed in Fig. 3.6. It can be shown (and verified on Fig. 3.6) that all these flux limiters should lie between the minmod and the Superbee limiters.

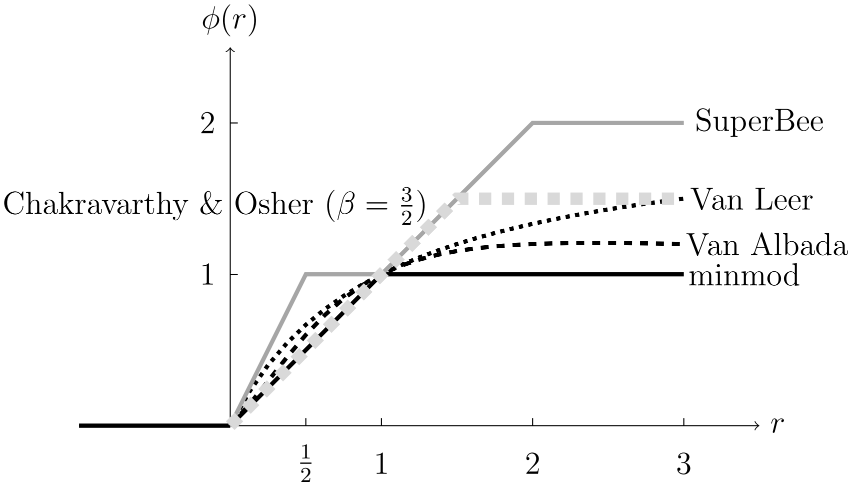

Figure 3.6: Representation of the flux limiter functions $\phi(r)$ for the minmod (solid black), Chakravarthy (wide dashed graa liney), Van Leer (dotted black line), Van Albada (dashed black line) and Superbee (solid gray line).

Property 11. *The minmod limiter is the most robust flux limiter, but then also the most dissipative.*

Property 12. *The Superbee limiter is the less robust flux limiter, but allows the stieffest shock fronts.*

Remark 42. *Obviously, $\phi(r) = 0$ gives the Roe-first order upwind scheme and $\phi(r) = 1$ gives the Lax-Wendroff scheme.*

3.3.4 TVD property and ENO

[Harten, 1983] introduced the concept of **Total Variation Diminishing**. This notion is very important in order to prevent the growth of any small-scale oscillations, generally close to stiff gradients.

For any function $u(x,t)$ of class C^1, its total variation is defined as

$$\mathrm{TV}[u,\tau] = \int_{\mathbb{R}} |\partial_x u(x,\tau)| \,\mathrm{d}x \tag{3.64}$$

Definition 37. *The Total Variation Diminishing (***TVD***) property means that for a given function $u(x,t)$, $\mathrm{TV}[u,\tau]$ is decreasing with time τ.*

Hence, if $\mathrm{TV}[u,t_2] \leq \mathrm{TV}[u,t_1]$ for $t_2 > t_1$, we say that $u(x,t)$ verifies the TVD property. The TVD property is then a way to prevent the birth of spurious oscillation, meaning that any solution verifying the TVD property should behave smoothly.

The total variation can also be defined in the same way for a discrete serie ; for the approximate solution w_j^n the total variation writes

$$\mathrm{TV}[w^n] = \sum_{j=0}^{N-1} |w_{j+1}^n - w_j^n| \tag{3.65}$$

so a discret solution will verify the TVD property if

$$\mathrm{TV}[w^{n+1}] \leq \mathrm{TV}[w^n] \tag{3.66}$$

Property 13. *A numerical scheme is TVD if the numerical solution verifies the TVD property given by Eq. (3.66)*

As a consequence, any local enhancement of a gradient of the solution w_j^n during its time evolution will be balanced by a (larger) decrease of such a gradient in a different location. Then, for any TVD method, spurious oscillations close to discontinuities can neither birth nor grow.

Property 14. *Any flux limited method verifies the TVD property.*

Definition 38. *The acronym* **ENO** *means* **Essentially Non Oscillatory***. A ENO scheme is a scheme which guarantees that no spurious oscillation will birth and grow close to stiff gradients of the solution w_j^n.*

Remark 43. *While their definitions are different, the TVD and ENO properties are not that far... Nonetheless, for historical reasons, some schemes are "TVD" and some others are "ENO".*

We will see later in this course that the ENO (and WENO) schemes are based on a "reconstruction-evolution" method (see the definition below).

3.4 Flux corrected schemes

The idea behind this class of method is not that far from the flux limited methods. In conservative form, the numerical solution depends on the numerical fluxes

$$w_j^{n+1} = w_j^n - \nu[F_{j+1/2}^n - F_{j-1/2}^n] \tag{3.67}$$

but then, the numerical flux is (generally) defined as

$$F_{j+1/2}^n = F_{j+1/2}^{(1)} + F_{j+1/2}^{(C)} \tag{3.68}$$

in which $F_{j+1/2}^{(1)}$ is a flux that is "corrected" by $F_{j+1/2}^{(C)}$. This corrected flux is defined, in several (but not all) cases as

$$F_{j+1/2}^{(C)} = \mathrm{d}_{j+1/2}^n(F_{j+1/2}^{(1)}, F_{j+1/2}^{(2)}) \tag{3.69}$$

where $\mathrm{d}_{j+1/2}^n$ is a function depending on the two fluxes $F_{j+1/2}^{(1)}$ and $F_{j+1/2}^{(2)}$. Up to now, this class of method really looks like a reformulation of flux limited methods...

Definition 39. *In flux limited methods, $\phi_{j+1/2}^n(r)$ generally depends on* **ratios** *of solutions or fluxes. In flux corrected methods $\mathrm{d}_{j+1/2}^n(f^1, f^2)$ generally depends on* **differences** *between solutions or fluxes.*

As a consequence, the function $\phi_{j+1/2}^n(r)$ in flux limiter depends on a single parameter, while $\mathrm{d}_{j+1/2}^n(F^1, F^2)$ depends on two (or eventually more) parameters.

3.4.1 The "Flux Corrected Transport" (FCT) method

This scheme proposed by [Boris and Book, 1973] is a blend between a first order upwind method and the Lax-Wendroff method. This method is still used now a day (but eventually in "legacy" codes).

We firstly introduce the **modified Boris-Book first-order upwind method** which flux si given by

$$\nu F_{j+1/2}^{\mathrm{Boris-Book}} = \nu F_{j+1/2}^{\mathrm{Lax-Wendroff}} - \frac{1}{8}(w_{j+1}^n - w_j^n) \tag{3.70}$$

so that, by developping the Lax-Wendroff flux, this flux can be written in artificial viscosity form as

$$\nu F_{j+1/2}^{\mathrm{Boris-Book}} = \frac{1}{2}\nu[f(w_{j+1}^n) - f(w_j^n)] - \frac{1}{2}\left[\left(\nu A_{j+1/2}^n\right)^2 + \frac{1}{4}\right](w_{j+1}^n - w_j^n) \tag{3.71}$$

We can apply the flux limited method to $F_{j+1/2}^{\mathrm{Boris-Book}}$ and use its difference with $F_{j+1/2}^{\mathrm{Lax-Wendroff}}$ in the limitor

$$F_{j+1/2}^n = F_{j+1/2}^{\mathrm{Boris-Book}} + \phi_{j+1/2}^n(F_{j+1/2}^{\mathrm{Lax-Wendroff}} - F_{j+1/2}^{\mathrm{Boris-Book}}) \tag{3.72}$$

and then work on the $\phi_{j+1/2}^n$ function.

Considering that shocks are the cause of spurious oscillations and extremas, we then would like to have $F^n_{j+1/2} \sim F^{\text{Boris-Book}}_{j+1/2}$ near extrema, and $F^n_{j+1/2} \sim F^{\text{Lax-Wendroff}}_{j+1/2}$ elsewhere. In term of corrective flux,

$$F^{(C)}_{j+1/2} = \begin{cases} 0 & \text{near extrema} \\ F^{\text{Lax-Wendroff}}_{j+1/2} - F^{\text{Boris-Book}}_{j+1/2} & \text{elsewhere} \end{cases} \tag{3.73}$$

which writes in term of flux limiters

$$\phi^n_{j+1/2} = \begin{cases} 0 & \text{near extrema} \\ 1 & \text{elsewhere} \end{cases} \tag{3.74}$$

Boris & Book also impose conditions on $F^n_{j+1/2}$, to stay between $F^{\text{Boris-Book}}_{j+1/2}$ and $F^{\text{Lax-Wendroff}}_{j+1/2}$, which is equivalent to $0 \leq \phi^n_{j+1/2} \leq 1$. A simple way to satisfy these 2 conditions is to take

$$\nu F^{(C)}_{j+1/2} = \text{minmod}\left[w^n_j - w^n_{j-1}, \frac{1}{8}(w^n_{j+1} - w^n_j), w^n_{j+2} - w^n_{j+1}\right] \tag{3.75}$$

which also writes

$$\phi^n_{j+1/2} = \text{minmod}\left(8r^+_j, 1, \frac{8}{r^+_j}\right) \tag{3.76}$$

with

$$r^+_j = \frac{w^n_j - w^n_{j-1}}{w^n_{j+1} - w^n_j} \tag{3.77}$$

Definition 40. *The minmod (minimum modulus) function equals the argument with the least absolute value if all the arguments have the same sign, and equal zero otherwise.*

This is the **One-step Boris-Book flux-corrected method**. But the most useful method is a two-setps variant of this one, that is a predictor

$$w^\star_j = w^n_j - \eta[F^{\text{Lax-Wendroff}}_{j+1/2} - F^{\text{Boris-Book}}_{j+1/2}] \tag{3.78}$$

followed by a corrector

$$w^{n+1}_j = w^\star_j - \eta[F^{(C)}_{j+1/2} - F^{(C)}_{j-1/2}] \tag{3.79}$$

with

$$\eta F^{(C)}_{j+1/2} = \text{minmod}(w^\star_j - w^\star_{j-1}, \frac{1}{8}[w^\star_{j+1} - w^\star_j], w^\star_{j+2} - w^\star_{j+1}) \tag{3.80}$$

This scheme is pretty good, both at shocks/discontinuities and regular solutions, provided that the CFL number is not too large (that is for low Mach number).

3.5 Reconstruction-Evolution method

The flux-averaged method can use numerical fluxes of the form $\frac{1}{2}[F_1(x) + F_2(x)]$. For solution-averaged methods, such a flux could instead be written as $f(\frac{1}{2}x_1 + \frac{1}{2}x_2)$, or more complicated (that is non-linear) combinations of x_1 and x_2, generaly resulting from polynomial interpolations. Such polynomial interpolation, (or "reconstruction") of the spatial form of the approximated solution is at the the heart of this class of methods.

3.5.1 The big picture

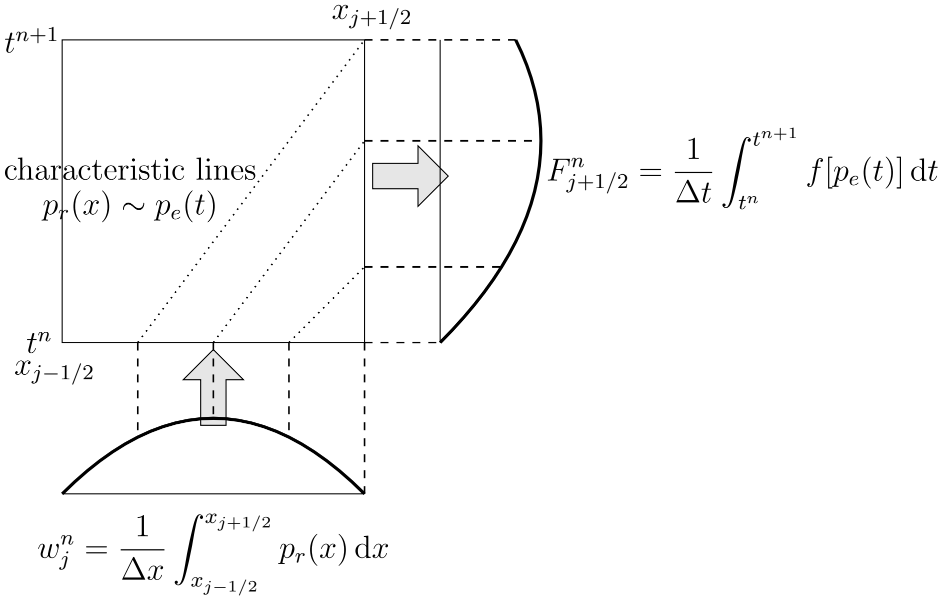
$x_{j+1/2}$
t^{n+1}
characteristic lines
$p_r(x) \sim p_e(t)$
$F^n_{j+1/2} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f[p_e(t)]\,\mathrm{d}t$
t^n
$x_{j-1/2}$
$w^n_j = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} p_r(x)\,\mathrm{d}x$

Figure 3.7: Basic principle of reconstruction-evolution methods.

The main idea behind **reconstruction-evolution** methods is essentially displayed in Fig. 3.7. Suppose that you have the $u(x_j, t^n)$. In each cell, you can build at t^n a polynom $p_r(x) \sim u(x, t^n)$, and an associated mean value

$$w^n_j = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} p_r(x)\,\mathrm{d}x \tag{3.81}$$

Notation 19. *$p_r(x)$ is a reconstructed polynom (at a given order) which approximates the true solution $u(x,t)$.*

Then, along a caracteristic, $p_r(x)$ is advected. Along a caracteristics, we have $p_r(x) = p_e(t)$ where $p_e(t) \sim u(x_{j+1/2}, t)$ is a time reconstruction of u at $x_{j+1/2}$.

Remark 44. *$p_r(x)$ will eventually reach $x_{j+1/2}$ (or $x_{j-1/2}$) before t^{n+1} as it is the case for $x_{j-1/2}$ in Fig. 3.7.*

Notation 20. *$p_e(t)$ is the evolution polynom (at a given order) that is the polynom which approximates $u(x_{j+1/2}, t)$ for $t^n < t < t^{n+1}$ (following the caracteristics).*

Finally, the flux can be reconstructed as

$$F^n_{j+1/2} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f[p_e(t)]\,\mathrm{d}x \tag{3.82}$$

For each of these steps, we have to define a form for $p_r(x)$ and $p_e(t)$. As a basic example, we can use a **piecewise spatial reconstruction** and a **exact temporal evolution**. We then have $p_r(x) = w^n_j$ and $p_e(t) = u[x_{j+1/2} - A(t - t^n), t^n]$. For a piecewise-constant spatial reconstruction, $p_e(t)$ depends on the sign of A, that is

$$p_e(t) = \begin{cases} w^n_j & \text{for } 0 \leq \nu A \leq 1 \\ w^n_{j+1} & \text{for } -1 \leq \nu A \leq 0 \end{cases} \tag{3.83}$$

meaning that the flux is then

$$F^n_{j+1/2} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f[p_e(t)]\,\mathrm{d}t = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} A p_e(t)\,\mathrm{d}t = \begin{cases} A w^n_j & \text{for } 0 \leq \nu A \leq 1 \\ A w^n_{j+1} & \text{for } -1 \leq \nu A \leq 0 \end{cases} \tag{3.84}$$

so that finally, the scheme writes

$$w^{n+1}_j = w^n_j - \nu A \begin{cases} w^n_j - w^n_{j-1} & \text{for } A > 0 \\ w^n_{j+1} - w^n_j & \text{for } A < 0 \end{cases} \tag{3.85}$$

Remark 45. *This kind of method is also called "Reconstruct-Solve-Average" :* **Reconstruct** *an approximate of $p_r(x)$,* **Solve** *the advection equation $p_r(x) = p_e(t)$, and* **Average** *to get the flux by Eq. (3.82).*

3.5.2 The Van Leer's method (MUSCL)

The MUSCL method is a class of methods that can be used with different choices. For the sake of simplification, we present the method by [Leer, 1977] in the simple case of a constant wave speed A. The non-linear case is more subtle. The MUSCL method uses a piecewise-linear spatial reconstruction and an exact temporal evolution. Then,

$$p_r(x) = w^n_j + S^n_j(x - x_j) \tag{3.86}$$

and

$$p_e(t) = u(x_{j+1/2} - A(t - t^n), t^n) \tag{3.87}$$

The slope S_j^n for the piecewise-linear polynom will be addressed later. Along a caracteristic, the exact temporal evolution gives

$$p_e(t) = p_r(x_{j+1/2} - A(t-t^n)) \tag{3.88}$$

Then, shifting the reconstruction to the left (or right, depending on the sign of A) by an amount At gives the form of $p_e(t)$

$$p_e(t) = \begin{cases} w_{j+1}^n - S_{j+1}^n[\frac{1}{2}\Delta x + A(t-t^n)] & \text{for } -1 \leq \nu A \leq 0 \\ w_j^n - S_j^n[\frac{1}{2}\Delta x - A(t-t^n)] & \text{for } 0 \leq \nu A \leq 1 \end{cases} \tag{3.89}$$

Focusing on the $0 \leq \nu A \leq 1$ case, we obtain the flux by average

$$F_{j+1/2}^n = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f[p_e(t)]\,\mathrm{d}t = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} Ap_e(t)\,\mathrm{d}t = \frac{A}{\Delta t}\int_{t^n}^{t^{n+1}} \left\{w_j^n + S_j^n\left[\frac{\Delta x}{2} - A(t-t^n)\right]\right\}\mathrm{d}t \tag{3.90}$$

Carying the integration gives

$$F_{j+1/2}^n = Aw_j^n + \frac{1}{2}A(1-\nu A)S_j^n\Delta x \tag{3.91}$$

Following the same process for $-1 \leq \nu A \leq 0$, we finally obtain the form of the flux

$$F_{j+1/2}^n = \begin{cases} Aw_j^n + \frac{1}{2}A(1-\nu A)S_j^n\Delta x & \text{for } 0 \leq \nu A \leq 1 \\ Aw_{j+1}^n - \frac{1}{2}A(1+\nu A)S_{j+1}^n\Delta x & \text{for } -1 \leq \nu A \leq 0 \end{cases} \tag{3.92}$$

It is of course not the end of the story because the slope S_j^n to use for the spatial reconstruction has to be defined. The temporal evolution being exact, the performance of this scheme then totally depends on the spatial reconstruction. Among the large variety of possibilities, the most obvious ones are

- the backward space approximation $S_j^{(\mathrm{BS})} = \dfrac{w_j^n - w_{j-1}^n}{\Delta x}$
- the forward space approximation $S_j^{(\mathrm{FS})} = \dfrac{w_{j+1}^n - w_j^n}{\Delta x}$
- the centered space approximation $S_j^{(\mathrm{CS})} = \dfrac{w_{j+1}^n - w_{j-1}^n}{2\Delta x}$

One constraint on the choice of S_j^n is to satisfy the upwind range condition,

$$|S_{j+1}^n - S_j^n| \leq 2\left|\frac{w_{j+1}^n - w_j^n}{\Delta x}\right| \tag{3.93}$$

In its original paper, Van Leer proposed

$$S_j^n = \mathrm{minmod}\left(2\frac{w_j^n - w_{j-1}^n}{\Delta x}, S_j^{(\mathrm{CS})}, 2\frac{w_{j+1}^n - w_j^n}{\Delta x}\right) \tag{3.94}$$

3.5.3 An example of ENO method

This one has been proposed by [Harten et al., 1987]. The MUSCL method uses an exact temporal evolution. For this second order ENO method, the temporal evolution phase is approximated by a second-order accurate time expansion. The definition of the numerical flux is still

$$F^n_{j+1/2} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f[u(x_{j+1/2}, t)]\,\mathrm{d}t \tag{3.95}$$

The integral can be avoided using a mid-point evaluation, that is

$$F^n_{j+1/2} = f[u(x_{j+1/2}, t^{n+1/2})] \tag{3.96}$$

As in the previous method, we still use a second-order accurate piecewise-linear reconstruction :

$$u(x, t^n) = w^n_j + S^n_j(x - x_j) \tag{3.97}$$

for $x_{j-1/2} \leq x \leq x_{j+1/2}$. In Eq. (3.96), we need the value of u forward in time, at $t^{n+1/2}$. To get this value at second order, we need a Taylor serie expansion

$$u(x,t) \sim u(x_j, t^n) + \partial_t u(x_j, t^n)(t - t^n) + \partial_x u(x_j, t^n)(x - x_j) \tag{3.98}$$

With Eq. (3.97) we then have directly

$$\partial_x u(x_j, t^n) \sim S^n_j \tag{3.99}$$

In the Taylor serie expansion of Eq. (3.98), there is also a temporal derivative. It can be evaluated because there is a direct link between spatial and temporal derivations, resulting from the structure of the linear advection equation. This so-called Cauchy-Kowalewsky procedure then gives

$$\partial_t u(x_j, t^n) = -A[u(x_j, t^n)]\partial_x u(x_j, t^n) \tag{3.100}$$

so that

$$\partial_t u(x_j, t^n) \sim -Aw^n_j S^n_j \tag{3.101}$$

The Taylor expansion (both in space and time) then writes

$$u(x,t) \sim u(x_j, t^n) - Aw^n_j S^n_j(t - t^n) + S^n_j(x - x_j) \tag{3.102}$$

so at the mid-point $x_{j+1/2}$

$$u(x_{j+1/2}, t^n) \sim u(x_j, t^n) + \frac{1}{2}[1 - \nu A w^n_j] S^n_j \Delta x \tag{3.103}$$

which is the reconstruction around the x_j point. Rewriting Eq. (3.97) around x_{j+1}

$$u(x, t^n) = w^n_{j+1} + S^n_{j+1}(x - x_{j+1}) \tag{3.104}$$

Eq. (3.103), can then also be written around point x_{j+1}

$$u(x_{j+1/2}, t^n) \sim u(x_{j+1}, t^n) - \frac{1}{2}[1 + \nu A w^n_{j+1}] S^n_{j+1} \Delta x \tag{3.105}$$

The forms of $u(x_{j+1/2}, t^n)$ given by Eq. (3.103) and (3.105) can be used to calculate the flux provided that we know which one to chose or how to "average" these values.

One option is to recall the Godunov average of Eq. (3.28)

$$F^n_{j+1/2} = \overline{F}^{\text{Godunov}}(w^n_j, w^n_{j+1}) = \begin{cases} \min\limits_{w^n_j \leq u \leq w^n_{j+1}} f(u) & \text{if } w^n_j < w^n_{j+1} \\ \max\limits_{w^n_j \geq u \geq w^n_{j+1}} f(u) & \text{if } w^n_j > w^n_{j+1} \end{cases} \tag{3.106}$$

We can then finally write the flux for the second-order ENO method as

$$F^n_{j+1/2} = \overline{F}^{\text{Godunov}}\left(w^n_j + \frac{1}{2}[1 - \nu A w^n_j] S^n_j \Delta x, w^n_{j+1} - \frac{1}{2}[1 + \nu A w^n_{j+1}] S^n_{j+1} \Delta x\right) \tag{3.107}$$

The main idea of ENO is to use a (space) polynomial reconstruction of $u(x, t^n)$ to be used for its temporal evolution. For a given width stencil (let say 3), the value at $x_{j+1/2}$ could use different set of points : $\{x_{j-2}, x_{j-1}, x_j\}$, $\{x_{j-1}, x_j, x_{j+1}\}$, $\{x_j, x_{j+1}, x_{j+2}\}$. In the ENO method described above, this choice was made using the Godunov's first-order upwind flux at grid points j and $j + 1$.

The ENO method can be improved if the grid points are not fixed but chosen deping on the regularity of the solution. The general idea to make this choice is to chose the stencil in which the solution will be the smoothest. Doing so, we make sure that the high order method are then limited to complex smooth solution while discontinuities do not suffer the instability and oscillations associted to Gibbs phenomena.

3.5.4 An example of WENO method

The WENO method is a "Weighted" ENO method, introduced by [Jiang and Shu, 1996]. The main iead is that instead of having to choose which one of the set of points is used for the polynomial reconstruction, the WENO method uses all of them, but with a "nonlinear weihgt", defined in the way that "good" set of points are significantly contributing while "bad" ones are not.

Consider a second order polynom for the interpolation of a set of w_j. Using the set (which is also a stencil) $S_2(j) = \{x_{j-2}, x_{j-1}, x_j\}$, we then have

$$w^{(2)}_{j+1/2} = \frac{1}{3} w_{j-2} - \frac{7}{6} w_{j-1} + \frac{11}{6} w_j \tag{3.108}$$

which is a third-order accuracy approximation because

$$w^{(2)}_{j+1/2} - u(x_{j+1/2}) = O(\Delta x^3) \tag{3.109}$$

Using the set $S_1(j) = \{x_{j-1}, x_j, x_{j+1}\}$, we then have

$$w^{(1)}_{j+1/2} = -\frac{1}{6}w_{j-1} + \frac{5}{6}w_j + \frac{1}{3}w_{j+1} \tag{3.110}$$

and using the set $S_0(j) = \{x_j, x_{j+1}, x_{j+2}\}$, we have

$$w^{(0)}_{j+1/2} = \frac{1}{3}w_j + \frac{5}{6}w_{j+1} - \frac{1}{6}w_{j+2} \tag{3.111}$$

The expressions of $w^{(r)}_{j+1/2}$ for $r = 0, 1, 2$ could be combined in a linear way to get a better approximation of $w_{j+1/2}$. We can verify that

$$w_{j+1/2} = \sum_{r=0}^{3} d_r w^{(r)}_{j+1/2} \tag{3.112}$$

is a fifth-order approximation of $w_{j+1/2}$ for the weights

$$d_2 = \frac{1}{10}\ ,\ d_1 = \frac{3}{5}\ ,\ d_0 = \frac{3}{10} \tag{3.113}$$

Such a linear combination is still a ENO scheme. Instead, the WENO method uses nonlinear weights to adaptively avoid to include the discontinuous cell in the stencil. Then, $w_{j+1/2}$ is a convex combination of the $w^{(r)}_{j+1/2}$'s

$$w_{j+1/2} = \sum_{r=0}^{3} \omega_r w^{(r)}_{j+1/2} \tag{3.114}$$

where the weights ω_r have to satisfy $\omega_r \geq 0$ and $\sum_{r=0}^{3} \omega_r = 1$. Jiang & Shu defined them as

$$\omega_r = \frac{\alpha_r}{\sum_{s=0}^{k-1} \alpha_s} \tag{3.115}$$

with $k = 3$ the order of accuracy of the $w^{(r)}_{j+1/2}$'s and the α_r's are defined as

$$\alpha_r = \frac{d_r}{(\varepsilon + \beta_r)^2} \tag{3.116}$$

The small parameter ε is generally taken to be 10^{-6} just in order to avoid the denominator to be null. The linear weights d_r's are given by Eq. (3.113). Following [Jiang and Shu, 1996], the smoothness indicator β_r are given by

$$\beta_r = \sum_{l=1}^{k-1} \Delta x^{2l-1} \int_{x_{j-1/2}}^{x_{j+1/2}} [d^l_{x^l} p_r(x)]^2 \, \mathrm{d}x \tag{3.117}$$

For the fifth-order WENO, the β_r's are then defined by

$$\beta_0 = \frac{13}{12}(w_j - 2w_{j+1} + w_{j+2})^2 + \frac{1}{4}(3w_j - 4w_{j+1} + w_{j+2})^2 \tag{3.118}$$

$$\beta_1 = \frac{13}{12}(w_{j-1} - 2w_j + w_{j+1})^2 + \frac{1}{4}(w_{j-1} - w_{j+1})^2 \tag{3.119}$$

$$\beta_2 = \frac{13}{12}(w_{j-2} - 2w_{j-1} + w_j)^2 + \frac{1}{4}(w_{j-2} - 4w_{j-1} + 3w_j)^2 \tag{3.120}$$

To build a finite volume scheme, we need to verify the upwinding property. Then, in Eq. (3.1), for the numerical flux $F_{j+1/2}$ we use a form

$$F_{j+1/2} = \overline{F}(w^-_{j+1/2}, w^+_{j+1/2}) \tag{3.121}$$

where $\overline{F}(a, b)$ is a monotone numerical flux. One such flux can be the Godunov flux given by Eq. (3.28), or the Lax-Friedrichs flux given by

$$\overline{F}^{\text{Lax}-\text{Friedrichs}}(a, b) = \frac{1}{2}[f(a) + f(b)] - \frac{1}{2}\alpha(b - a) \tag{3.122}$$

where $\alpha = \max_u |f'(u)|$ is a constant value calculated on the appropriate range of u. In this expression, $w^-_{j+1/2}$ and $w^+_{j+1/2}$ are WENO approximations based on cell average values in stencils one-cell biased to the left and one-cell biased to the right, respectively. For a fifth-order WENO scheme, the value of $w^-_{j+1/2}$ uses the cell points $\{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$ (and is the one we presented hereabove) while $w^+_{j+1/2}$ uses the cell points $\{x_{j-1}, x_j, x_{j+1}, x_{j+2}, x_{j+3}\}$ (which can be reconstructed in the very same way).

Remark 46. *With such nonlinear weights, the WENO approximation is fifth-order accurate if the function $u(x)$ is smooth in the large stencil $S_2(j) \cup S_1(j) \cup S_0(j)$. If $u(x)$ is not smooth in a stencil $S_r(j)$, but is smooth in at least one of the two other stencils, then the WENO scheme guarantee a non-oscillatory result since the contribution from any stencil containing the discontinuity of $u(x)$ has an essentially zero weight.*

Bibliography

[Beam and Warming, 1976] Beam, R. M. and Warming, R. (1976). An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *Journal of Computational Physics*, 22(1):87–110.

[Boris and Book, 1973] Boris, J. P. and Book, D. L. (1973). Flux-corrected transport. i. shasta, a fluid transport algorithm that works. *Journal of Computational Physics*, 11(1):38–69.

[Courant et al., 1928] Courant, R., Friedrichs, K., and Lewy, H. (1928). Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74.

[Courant et al., 1952] Courant, R., Isaacson, E., and Rees, M. (1952). On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics*, 5(3):243–255.

[Fromm, 1968] Fromm, J. E. (1968). A method for reducing dispersion in convective difference schemes. *Journal of Computational Physics*, 3(2):176–189.

[Godlewski and Raviart, 1996] Godlewski, E. and Raviart, P.-A. R. (1996). *Numerical approximation of hyperbolic systems of conservation laws.* Applied mathematical sciences ; v. 118. Springer, New York.

[Godunov, 1959] Godunov, S. K. (1959). Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematičeskij sbornik*, 47(89)(3):271–306.

[Harten, 1983] Harten, A. (1983). High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–393.

[Harten et al., 1987] Harten, A., Engquist, B., Osher, S., and Chakravarthy, S. R. (1987). Uniformly high order accurate essentially non-oscillatory schemes, III. *Journal of Computational Physics*, 71(2):231–303.

[Jiang and Shu, 1996] Jiang, G.-S. and Shu, C.-W. (1996). Efficient implementation of weighted eno schemes. *Journal of Computational Physics*, 126(1):202–228.

[Laney, 1998] Laney, C. B. (1998). *Computational Gasdynamics.* Cambridge University Press.

[Lax and Wendroff, 1960] Lax, P. and Wendroff, B. (1960). Systems of conservation laws. *Communications on Pure and Applied Mathematics*, 13(2):217–237.

[Lax, 1954] Lax, P. D. (1954). Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications on Pure and Applied Mathematics*, 7(1):159–193.

[Leer, 1977] Leer, B. V. (1977). Towards the ultimate conservative difference scheme. IV. a new approach to numerical convection. *Journal of Computational Physics*, 23(3):276–299.

[macCormack, 1969] macCormack, R. (1969). The effect of viscosity in hypervelocity impact cratering. In *4th Aerodynamic Testing Conference*. American Institute of Aeronautics and Astronautics.

[Morton and Mayers, 1994] Morton, K. W. and Mayers, D. F. (1994). *Numerical solution of partial differential equations*. Cambridge University Press.

[Richtmyer, 1962] Richtmyer, R. (1962). A survey of difference methods for non-steady fluid dynamics. Technical report.

[Roe, 1981] Roe, P. (1981). Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357–372.

[Sweby, 1999] Sweby, P. (1999). *Godunov Methods*. Numerical analysis report. University of Reading, Department of Mathematics.

[Sweby, 1984] Sweby, P. K. (1984). High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal on Numerical Analysis*, 21(5):995–1011.

[Toro, 2009] Toro, E. F. (2009). *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer Berlin Heidelberg.

[van Leer, 1974] van Leer, B. (1974). Towards the ultimate conservative difference scheme. II. monotonicity and conservation combined in a second-order scheme. *Journal of Computational Physics*, 14(4):361–370.