Université Libre de Bruxelles

Summary

Numerical methods in aerothermodynamics MECA-H-407

Author:
Enes Ulusoy

Professor:
Gérard Degrez

Year 2016 - 2017

Appel à contribution

Synthèse Open Source

Ce document est grandement inspiré de l'excellent cours donné par Gérard Degrez à l'EPB (École Polytechnique de Bruxelles), faculté de l'ULB (Université Libre de Bruxelles). Il est écrit par les auteurs susnommés avec l'aide de tous les autres étudiants et votre aide est la bienvenue ! En effet, il y a toujours moyen de l'améliorer surtout que si le cours change, la synthèse doit être changée en conséquence. On peut retrouver le code source à l'adresse suivante

`https://github.com/nenglebert/Syntheses`

Pour contribuer à cette synthèse, il vous suffira de créer un compte sur *Github.com*. De légères modifications (petites coquilles, orthographe, ...) peuvent directement être faites sur le site ! Vous avez vu une petite faute ? Si oui, la corriger de cette façon ne prendra que quelques secondes, une bonne raison de le faire !

Pour de plus longues modifications, il est intéressant de disposer des fichiers : il vous faudra pour cela installer LaTeX, mais aussi *git*. Si cela pose problème, nous sommes évidemment ouverts à des contributeurs envoyant leur changement par mail ou n'importe quel autre moyen.

Le lien donné ci-dessus contient aussi un `README` contenant de plus amples informations, vous êtes invités à le lire si vous voulez faire avancer ce projet !

Licence Creative Commons

Le contenu de ce document est sous la licence Creative Commons : *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)*. Celle-ci vous autorise à l'exploiter pleinement, compte- tenu de trois choses :

1. *Attribution* ; si vous utilisez/modifiez ce document vous devez signaler le(s) nom(s) de(s) auteur(s).
2. *Non Commercial* ; interdiction de tirer un profit commercial de l'œuvre sans autorisation de l'auteur
3. *Share alike* ; partage de l'œuvre, avec obligation de rediffuser selon la même licence ou une licence similaire

Si vous voulez en savoir plus sur cette licence :

`http://creativecommons.org/licenses/by-nc-sa/4.0/`

Merci !

Contents

Introduction

Fluid dynamics is based on continuity hypothesis, all quantities can be expressed as a continuous function of time and space coordinates. The governing equations are partial differential equations. Because of the geometrical complexity of the domain and of the equations, we need strategies. The first one is to forget about the equations and to rely on experiments. The second is to consider simplified cases, and approximate theoretical analytic solutions (aerodynamics). The third approach is numerical approach. Disadvantages and advantages of the different methods can be listed as:

- **Experimental:** the advantage is that it is the most realistic, but requires equipment, there are scale problems (similarity), interferences (tests in finite space), measurement difficulties and operating costs.
- **Theoretical**: the advantage is that we have a mathematical expression and we don't have to repeat calculus, but it is restricted to simple geometries and linear problems.
- **Numerical**: the advantages are that we can dead with complex geometries, non linear problems and unsteady problems, but there are truncation errors, problems with boundary conditions like the finite space in experiment and the computation cost.

In reality these approaches are complementary. We can use the second method to simplify the numerical computations, crucial for example for costly operations like computations on turbulence. The evolution of numerical cost over the past 40 years has been particularly impressive, cost decreased dramatically. In the other hand, the experimental cost tends to increase (technical personal, material,...). Nowadays we can measure many things impossible to measure before. This explains why the numerical computations have spread incredibly.

The design relies mainly on the numerical methods and less on the experimental testing, but it is still needed to confirm the data. We can use the numerical methods in many fields and we could call this « numerical physics ». We should deal with this in a single course of computational method and then to specialize it to the specific disciplines. An approximate solution to a problem is some kind of mathematical entity, an object, which depends on a finite number of real parameters, and which constitutes a representation of the continuous field under study. The numerical solution belongs to a finite dimension space whereas the theoretical to an infinite.

There are different types of numerical representation:

- **Discrete:** collection of either point values (samples of the solutions) or subdomain averages. We are not able to give an exact solution on basis of these points, but rather an estimation.
- **Functional:** the solution is expressed as a function $u*(x) = f(x, a_i)$, depending on a set of variables. Most of the time the dependence is linear: $u* = \sum_{i=1} na_i v_i(x)$ where $v_i(x)$ are a priori specified functions.

Ones we have chosen the numerical representation method, we have to generate a system of algebraic equations linking the parameters from the representation and the governing equations. The last step is to solve the system. The step of generating the equations system is called **discretization**. For one problem, several discretizations are possible. Some examples are given, consult the syllabus for more details.

Chapter 1

Discretization methods

1.1 Finite difference method

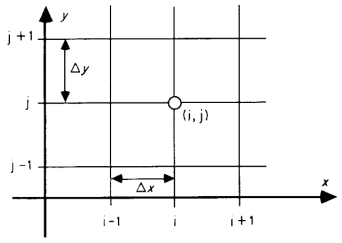

Figure 1.1

It is based on discrete representation of numerical solution consisting of values of the solution at the nodes of a Cartesian mesh of uniform spacing (Figure 1.1). The discretization will consist in estimating the partial derivatives appearing in the governing equations by some algebraic relations at the nodes. Because of the limitation of the mesh we cannot deal with curved geometries. To do so we need coordinate changes. In this type of mesh, each grid point can be identified by a set of indexes i, j, (k), this type of mesh is called **structured mesh**. The neighbors are implicitly given by the index identification.

1.1.1 Evaluation of derivatives by finite differences

Difference formulas for the derivative $\partial u/\partial x$

The value of a function $u(x)$ on a point of indexes i, j on the mesh is noted u_{ij}, for time dependent functions we use u_i^n where n denotes the time index and i the space index. Let's estimate $(\partial u/\partial x)_{ij}$, by definition:

$$\left.\frac{\partial u}{\partial x}\right|_{ij} = \left.\frac{\partial u}{\partial x}\right|_{x_0,y_0} = \lim_{\xi\to 0}\frac{u(x_0+\xi,y_0)-u(x_0,y_0)}{\xi} \tag{1.1}$$

By taking $\xi = \Delta x$ to fit our grid points we get:

$$\left.\frac{\partial u}{\partial x}\right|_{ij} \approx \lim_{\Delta x\to 0}\frac{u(x_0+\Delta x,y_0)-u(x_0,y_0)}{\Delta x} = \frac{u_{i+1\,j}-u_{ij}}{\Delta x} \tag{1.2}$$

In order to find a systematic way of deriving the equations, let's build the Taylor expansion of $u(x,y)$ around the mesh point ij:

$$\begin{aligned} u_{i+1\,j} &= u_{ij} + \Delta x\left.\frac{\partial u}{\partial x}\right|_{ij} + \frac{\Delta x^2}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_{ij} + \frac{\Delta x^3}{6}\left.\frac{\partial^3 u}{\partial x^3}\right|_{\xi} \qquad x_0\le\xi\le x_0+\Delta x \\ \Leftrightarrow \left.\frac{\partial u}{\partial x}\right|_{ij} &= \frac{u_{i+1\,j}-u_{ij}}{\Delta x} \underbrace{- \frac{\Delta x}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_{ij} - \frac{\Delta x^2}{6}\left.\frac{\partial^3 u}{\partial x^3}\right|_{\xi}}_{\text{truncation error}} = \frac{u_{i+1\,j}-u_{ij}}{\Delta x} + \mathcal{O}(\Delta x) \end{aligned} \tag{1.3}$$

Giving the **forward difference formula**. We have then a truncation error whose behavior is dominated by the first term when $\Delta x \to 0$, so that $TE = \mathcal{O}(\Delta x)$, meaning that there exists a bounded number K such that $\Delta x < \epsilon \to |TE| < K\Delta x$. The truncation error is always in the form $TE = \mathcal{O}(\Delta x^q)$ where q is the order of accuracy. The forward finite difference approximation of $\frac{\partial u}{\partial x}$ is first order accurate since $q = 1$.

If the order of the method is larger, for example second order, it means that if $\Delta x \to 0$, after a certain Δx_{crit} the truncation error goes to 0 faster than the truncation error of a lower order method. If the mesh is not finer than a critical value, this is not true. When we go higher than a second order it is not clear in practice if we have something better because increasing the order allows the use of larger mesh, but is computationally more expensive too.

The definition for the derivative is not unique, for instance the **backward difference formula**:

$$u_{i-1\,j} = u_{ij} - \Delta x \left.\frac{\partial u}{\partial x}\right|_{ij} + \left.\frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\right|_{ij} + H.O.T \Leftrightarrow \left.\frac{\partial u}{\partial x}\right|_{ij} = \frac{u_{ij} - u_{i-1\,j}}{\Delta x} + \mathcal{O}(\Delta x) \tag{1.4}$$

We have an infinity of finite difference formula if make the linear combination of the two last expressions. For example, if we sum half of the two we get the **central finite difference formula**:

$$\left.\frac{\partial u}{\partial x}\right|_{ij} = \frac{u_{i+1\,j} - u_{i-1\,j}}{2\Delta x} + \mathcal{O}(\Delta x^2) \tag{1.5}$$

We see that the central difference formula is more accurate than the others and involves the same mesh distance. We could thus get as high order as desired, but at the cost of increasing the number of neighboring grid points in the equation and thus the computational cost.

General method to obtain finite difference formulas

- Choose the stencil (set of points involved in the expression);
- write Taylor series expansion of all the points in the stencil around the point where the derivative is to be evaluated;
- write the finite difference formula as a linear combination of stencil point values and adjust the coefficients such that it approximates the derivative to be evaluated with the desired order of accuracy.

EXAMPLE: Let's compute the finite difference formula for $\partial^2 u/\partial x^2$ using $i-1\,j, ij$ and $i+1\,j$ (first step). The second step gives:

$$\begin{aligned}
u_{i+1\,j} &= u_{ij} + \Delta x \left.\frac{\partial u}{\partial x}\right|_{ij} + \left.\frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\right|_{ij} + \left.\frac{\Delta x^3}{6}\frac{\partial^3 u}{\partial x^3}\right|_{ij} + \left.\frac{\Delta x^4}{24}\frac{\partial^4 u}{\partial x^4}\right|_{ij} + H.O.T \\
u_{ij} &= u_{ij} \\
u_{i-1\,j} &= u_{ij} - \Delta x \left.\frac{\partial u}{\partial x}\right|_{ij} + \left.\frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\right|_{ij} - \left.\frac{\Delta x^3}{6}\frac{\partial^3 u}{\partial x^3}\right|_{ij} + \left.\frac{\Delta x^4}{24}\frac{\partial^4 u}{\partial x^4}\right|_{ij} + H.O.T
\end{aligned} \tag{1.6}$$

The third step gives:

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{ij} = a u_{i+1\,j} + b u_{ij} + c u_{i-1\,j} = (a+b+c)u_{ij} + (a-c)\Delta x \left.\frac{\partial u}{\partial x}\right|_{ij} + (a+c)\left.\frac{\partial^2 u}{\partial x^2}\right|_{ij} + H.O.T \tag{1.7}$$

Depending on the accuracy needed, establish a system of 3 equations of 3 variables and solve by imposing the value for the different terms. For example if we want to approximate exactly we should cancel all the terms except the second order derivative term.

We can repeat this method again and again to obtain various finite difference formulas. You can consult pages 14 and 15 of the syllabus to see the list, not useful, just know that we can express the mixed second derivatives like $\partial^2 u/\partial x \partial y$ too.

Derivation of finite difference formulas using operators

In order to make the writing more compact, let's introduce some operators:

$E_x^{+1} u_{ij} = u_{i+1\,j}$	Forward shift	$E_x^{-1} u_{ij} = u_{i-1\,j}$	Backward shift
$\delta_x^{+} u_{ij} = u_{i+1\,j} - u_{ij}$	Forward difference	$\delta_x^{-} u_{ij} = u_{ij} - u_{i-1\,j}$	Backward difference
$\mu_x u_{ij} = \frac{1}{2}\left(u_{i+\frac{1}{2}\,j} + u_{i-\frac{1}{2}\,j}\right)$	Averaging	$\delta_x u_{ij} = u_{i+\frac{1}{2}\,j} - u_{i-\frac{1}{2}\,j}$	Centered difference

Another operator for the centered difference can be used:

$$\bar{\delta}_x = \frac{1}{2}\left(\delta_x^{+} + \delta_x^{-}\right) \qquad \Rightarrow \bar{\delta}_x u_{ij} = \frac{1}{2}\left(u_{i+1\,j} - u_{i-1\,j}\right) \tag{1.8}$$

All these operators are also valid for y coordinate. The following relations are verified:

$$\delta^{+} = E^{+1} - 1 \qquad \delta^{-} = 1 - E^{-1} \qquad \bar{\delta} = \mu\delta = \delta\mu \tag{1.9}$$

It is easy to derive finite difference formulas with these operator. For example, the Taylor series expansion of $u(x)$ is:

$$\begin{gathered} u(x + \Delta x) = u(x) + \Delta x \frac{\partial u}{\partial x}(x) + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}(x) + \dots \\ \Leftrightarrow Eu(x) = \left(1 + \Delta x D + \frac{(\Delta x D)^2}{2} + \dots\right) u(x) \qquad \Leftrightarrow Eu(x) = \exp(\Delta x D) u(x) \end{gathered} \tag{1.10}$$

where we clearly see the Taylor expansion of $\exp(\Delta x D)$ and where $D_x = \frac{\partial}{\partial x}$. We can then make the following manipulation:

$$E = \exp(\Delta x \partial) \leftrightarrow \ln(E + \mathbf{1} - \mathbf{1}) = \ln(1 + \delta^{+}) = \Delta x D \qquad \Rightarrow D = \frac{\ln(1 + \delta^{+})}{\Delta x} \tag{1.11}$$

And finally if we make the Mac Laurin expansion of the logarithm:

$$D = \frac{\delta^{+}}{\Delta x} - \frac{{\delta^{+}}^2}{2\Delta x} + \frac{{\delta^{+}}^3}{3\Delta x} + \dots \tag{1.12}$$

By keeping the first term we find the first order forward difference formula, by keeping the first two terms we find the second order one, and so on.

1.1.2 Finite difference formulas for partial differential equations

There are two strategies to express equations:

- **Strategy 1**: simply assemble the finite difference formula for each individual derivative;
- **Strategy 2**: same strategy used to find the finite difference in many steps, select the stencil, Taylor expansion on each point of the stencil, write the FD formula as a linear combination of the stencil points values and select the coefficients.

These methods can't be differentiated by using the trucation error. The first method is the most used. For the Laplace equation $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ we have:

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{ij} = \frac{{\delta_x}^2 u_{ij}}{\Delta x^2} + \mathcal{O}(\Delta x^2) \qquad \left.\frac{\partial^2 u}{\partial y^2}\right|_{ij} = \frac{{\delta_y}^2 u_{ij}}{\Delta y^2} + \mathcal{O}(\Delta y^2) \tag{1.13}$$

If we sum this up we get:

$$\frac{{\delta_x}^2 u_{ij}}{\Delta x^2} + \frac{{\delta_y}^2 u_{ij}}{\Delta y^2} = 0 \qquad \Rightarrow \frac{u_{i+1\,j} - 2u_{ij} + u_{i-1\,j}}{\Delta x^2} + \frac{u_{i\,j+1} - 2u_{ij} + u_{i\,j-1}}{\Delta y^2} = 0 \tag{1.14}$$

The equation can contain a first order derivative and there can thus exists several discretization (forward, backward, ...).

1.1.3 Arbitrary geometries - irregular meshes

The method we have seen is very simple, we love it. But the expressions rapidly become very difficult when dealing with irregular meshes. In addition, the order of accuracy is lower when irregular meshes compared to the regular one with same size mesh. The formulas become intractable for more than 3 points. We cannot only use uniform meshes for at least two reasons:

- **Computational domain geometry:** when the boundary is curved, it is quasi impossible to use uniform rectangular mesh. On the aerofoil example below, one can see that the grid points not always intersect the nodes on the geometry.
- **Presence of regions where the solution varies rapidly:** for example, in fluid mechanics, there are regions such as the boundary layer where the fluid properties vary more rapidly than anywhere else. It is thus interesting to have finer mesh there and larger mesh somewhere where we don't care.

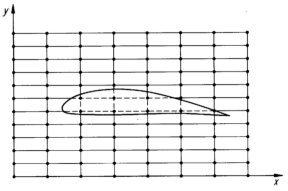

Figure 1.2

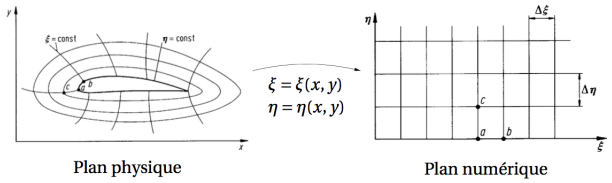

Figure 1.3

To tackle these problems, one can use coordinate transformation as suggests Figure 1.3. One can thus first fit a certain geometry, but also achieve a local concentration of mesh points. There are two disadvantages to this: we transform the geometrical complexity into equation complexity, and it is very difficult to find these transformation (numerical methods needed). The good news are that in the numerical plane, the mesh is regular and numerical algorithms have high efficiency, the transformation will be discussed later.

1.2 Finite volume method

The main idea is to take advantage of conservation equations whose fundamental form is the **integral form**, we discretize the integral. The principle consist in the application of the control volume method (macroscopic balances) in a large scale. The classical example is to have a bent tube, the flow exerts a force on the elbow and we can easily estimate it by momentum balance.

We just take several small volumes where to apply this.

The great advantage of this method is to use arbitrary polygons (2D) or polyhedra (3D) as control volume so that it offers great **flexibility**. Unlike the finite difference method, the finite volume method can accommodate arbitrary control volume shapes, this eases mesh generation dramatically. There are some independent variables (time and space) that does not need flexibility it is the time variable, this is why wee still use finite differences for time discretization. In addition, since the integral form is discretized, it allows the computation of **weak solutions** of the flow.

1.2.1 Fundamental principles and variants of the method

Let's consider the integral form of a general system of conservation equations:

$$U = \begin{pmatrix} \rho \\ \rho\vec{u} \\ \rho E \end{pmatrix} \qquad \Rightarrow \frac{\partial U}{\partial t} + \nabla.\vec{F} = Q, \tag{1.15}$$

where $\vec{F}$ is the **flux vector** and Q the **source term**. If we take the momentum equation and the conservation equation, we can see that there is a part independent of the derivative of U (convective term) and a diffusive term dependent of $\nabla U \rightarrow \vec{F} = \vec{F}(U, \nabla U)$:

$$\frac{\partial \rho\vec{u}}{\partial t} + \nabla(\underbrace{\rho\vec{u} \otimes \vec{u} + p\bar{\bar{1}}}_{\text{convective}} - \underbrace{\bar{\bar{\tau}}}_{\text{diffusive}}) = \rho\vec{g} \tag{1.16}$$

The corresponding integral form is the basic original form obtained by integration of the equation over a control volume Ω:

$$\frac{d}{dt}\int_\Omega U d\Omega + \oint_{\partial\Omega} \vec{F}.\vec{n}\, dS = \int_\Omega Q\, d\Omega \tag{1.17}$$

Remark that discontinuities are allowed in this integral form since we do not have to verify the differentiation everywhere in the domain. If we subdivide the domain in elementary volumes and use the average value of U on that volumes $\int_{\Omega_k} U\, d\Omega = U_k\Omega_k$ these are chosen as the parameters of the discrete representation, and assume the control volume to be a polygon (Γ_m the faces), we have:

$$\frac{d}{dt}(\Omega_k U_k) + \sum_{\Gamma_m \in \partial\Omega_k} \int_{\Gamma_m} \vec{F}.\vec{n}\, dS = \int_{\Omega_k} Q\, d\Omega \tag{1.18}$$

To make the discretization, we need to evaluate the remaining surface and volume integrals in terms of neighboring control volume averages. How to build the control volumes? First a mesh of non-overlapping elementary surfaces/volumes is generated, these are called cells. The design of control volumes must respect a certain number of conditions:

- the union of CVs must cover the whole domain of interest;
- the CVs may overlap but the boundaries of a CV should be either lying on the domain boundary or belong to the boundary of another CV. Each CV boundary must be shared by two CVs;
- the expression of the flux integral on a common edge should be the same for the two CVs it belongs to.

Consider two CVs K and L with a common face Γ_c and make the sum:

$$\begin{aligned} \frac{d}{dt}(\Omega_K U_K) + \sum_{\Gamma_m \in \partial\Omega_K} \int_{\Gamma_m} \vec{F}.\vec{n}\, dS &= \int_{\Omega_K} Q\, d\Omega \\ \frac{d}{dt}(\Omega_L U_L) + \sum_{\Gamma_m \in \partial\Omega_L} \int_{\Gamma_m} \vec{F}.\vec{n}\, dS &= \int_{\Omega_L} Q\, d\Omega \\ \Rightarrow \frac{d}{dt}(\Omega_K U_K + \Omega_L U_L) + \sum_{\Gamma_m \in \partial\Omega_K \cup \partial\Omega_L \setminus \Gamma_c} \int_{\Gamma_m} \vec{F}.\vec{n}\, dS &= \int_{\Omega_K \cup \Omega_L} Q\, d\Omega \end{aligned} \tag{1.19}$$

where we can observe that the common boundary integral disappears since the flux should be the same but the normals are opposite to each others. This last property is called **telescopic property** that ensures the conservation at the discrete level and the capture of discontinuities. Indeed, if the flux was different on K and L, the term would remain.

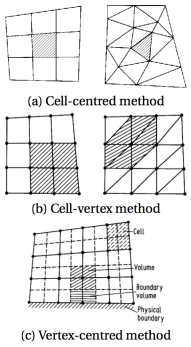

Figure 1.4

Several arrangement methods exists:

- the CV coincide with the mesh cell → cell-centered method;
- the CV is made out of mesh cells having a common vertex → cell-vertex method;
- the CV is made out of part of mesh cells sharing a common vertex → vertex centered method.

In the two last ones, it is common to associate the volume average to the corresponding vertex as done in finite difference. Only these will be considered. Let's mention that it is not compulsory to use the same CVs for different equations of a system of equations.

1.2.2 Evaluation of fluxes through faces

In general we will approximate the integral over a length/surface $Gamma_i$ by a one point quadrature integration formula:

$$\int_{\Gamma_m} \vec{F}.\vec{n}\, dS \approx \vec{F}_m.\vec{n}_m S_m \tag{1.20}$$

This is sufficient for first order and second order methods, for higher order you have to use more points quadrature. Moreover, higher order are not easy to construct, this is why we have finite element methods. Let's discuss about aerothermodynamic problems ($\vec{F} = \vec{F}(U, \nabla U)$) in 1D for simplicity. We will consider a vertex-centered 1D FV method and the CVs are segments. The discretization becomes:

$$\frac{d}{dt}(\Delta x_i U_i) + F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = 0 \tag{1.21}$$

How to express the $F_{i+1/2}$ and the other in function of cell averages? One needs to specify a **numerical flux formula** which plays the same role as finite difference formulas in finite difference method. We will say that:

$$F_{i+\frac{1}{2}} \approx \Phi(U_{i-k+1}, \ldots, U_{i+k}) \tag{1.22}$$

For the method to be at least of order one, the approximation should be exact for a uniform field $\Phi(U, \ldots, U) = F(U)$. The simplest choice is to take an arithmetic average of the fluxes or of the variables:

$$\Phi(U_i, U_{i+1}) = (F_i + F_{i+1})/2 \qquad \Phi(U_i, U_{i+1}) = F\left(\frac{U_i + U_{i+1}}{2}\right) \tag{1.23}$$

This applied to (1.21) gives:

$$\frac{d}{dt}(\Delta x_i U_i) + \frac{F_i + F_{i+1}}{2} - \frac{F_{i-1} + F_i}{2} = 0 \qquad \Rightarrow \frac{dU_i}{dt} + \frac{F_{i+1} - F_{i-1}}{2\Delta x_i} = 0 \tag{1.24}$$

which is the same expression as obtained by central finite difference formula. One can retrieve the first order forward and backward finite difference formula by choosing $\Phi(U_i, U_{i+1}) = F(U_{i+1})$ and $\Phi(U_i, U_{i+1}) = F(U_i)$.

Let's come back to the nature of the finite volume numerical representation. It consists of a set of average values over subdomains and is thus clearly a discrete representation. For cell-centered or vertex-centered methods (not overlapping) it is easy to reconstruct a functional representation out of the averages. A piecewise constant reconstruction and a linear reconstruction are illustrated below.

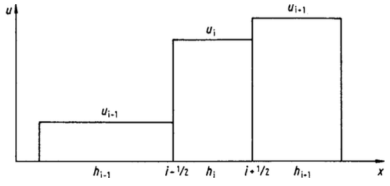

Figure 1.5

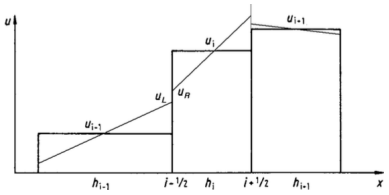

Figure 1.6

In the second, the solution gradient is estimated in each CV but the reconstruction remains discontinuous at the boundaries so that it does not eliminate the need for a numerical flux function to compute the flux across the boundaries. But it allows to easily construct more accurate flux functions, starting from a two variable flux function:

$$F_{i+\frac{1}{2}} \approx \Phi(U_i, U_{i+1}) \tag{1.25}$$

associated with the constant reconstruction, one obtains more accuracy by replacing U_i and U_{i+1} by U_L and U_R. For instance using the backward flux formula $\Phi(U_L, U_R) = F(U_L)$ and a gradient estimation in CV i based on the back point:

$$\left(\frac{\partial U}{\partial x}\right)_i \approx \frac{U_i - U_{i-1}}{\Delta x} \qquad \Rightarrow U_{L,i+\frac{1}{2}} = \frac{3}{2}U_i - \frac{1}{2}U_{i-1}, \tag{1.26}$$

One can obtain the following space discretization:

$$\frac{dU_i}{dt} + \frac{1}{\Delta x}\left(F\left(\frac{3U_i - U_{i-1}}{2}\right) - F\left(\frac{3U_{i-1} - U_{i-2}}{2}\right)\right) = 0 \tag{1.27}$$

and for a particular $F(U) = au$, we have:

$$\frac{dU_i}{dt} + a\frac{3U_i - 4U_{i-1} + U_{i-2}}{2\Delta x} \tag{1.28}$$

which is the one we found in previous section. Generally, for a polynomial reconstruction of order k we shall obtain a discretization of order at least $k + 1$. For the diffusive fluxes we have to estimate the gradient of variables on the faces which can be done directly or by averaging the estimated gradients in the two neighboring CVs. This is done by Green-Gauss theorem:

$$\int_\Omega \nabla U \, d\Omega = \oint_\Gamma U \vec{n} \, dS \tag{1.29}$$

by choosing an auxiliary control volume Ω centered on the point where one wishes to estimate the gradient.

1.2.3 Vertex-centered finite volumes in two dimensions and comparison with finite differences in transformed coordinates

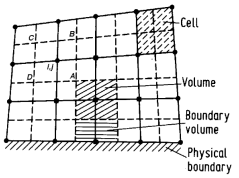

Figure 1.7

Here we will show that the FV method is equivalent to the FD in transformed coordinates. We consider first a vertex centered FV method on a structured mesh and then FD method in transformed coordinates. The advantage of structured mesh is the explicit connectivity, but its generation is more complex. We take the finite volume around the point i, j. Let's write the Euler equation:

$$\frac{\partial U}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} = \frac{\partial U}{\partial t} + \nabla . \vec{F} = 0 \tag{1.30}$$

The finite volume discretization over the 4 points can be written as:

$$\Omega_{ABCD} \frac{dU_{ij}}{dt} + \sum_{m=1}^{4} \vec{F}_m \vec{n}_m l_m = 0 \tag{1.31}$$

where $\vec{F}_m$ can be chosen as the average value (middle of sides) and $n_x l = \Delta y, n_y l = -\Delta x$ since if we consider the angle θ made by AB we have that $\vec{n} = \cos\theta \vec{e}_1 + \sin\theta \vec{e}_2$ with $\cos\theta = \frac{\Delta y_{AB}}{AB}$ and $\sin\theta = -\frac{\Delta x_{AB}}{AB}$. Ω_{ABCD} can be computed as the vector product of diagonals:

$$\Omega_{ABCD} = \frac{1}{2} |\Delta \vec{x}_{AC} \times \Delta \vec{x}_{BD}| = \frac{1}{2} (\Delta x_{AC} \Delta y_{BD} - \Delta x_{BD} \Delta y_{AC}) \tag{1.32}$$

We see that the remaining work to discretize is to express $\frac{dU_{ij}}{dt}$ by a finite difference formula. We could do in other way.

Let's now examine the transformed coordinates FD, let's do the chain rule and replace the expressions in Euler equation to get the transformed coordinates equation:

$$\frac{\partial F_x}{\partial x} = \xi_x \frac{\partial F_x}{\partial \xi} + \eta_x \frac{\partial F_x}{\partial \eta} \qquad \frac{\partial F_y}{\partial y} = \xi_y \frac{\partial F_y}{\partial \xi} + \eta_y \frac{\partial F_y}{\partial \eta} \tag{1.33}$$

$\xi_x, \xi_y, \eta_x \eta_y$ are the **metric terms**. We have the relation $\frac{dx}{d\xi}\frac{d\xi}{dx} = 1$ that can be generalized into matrix as follows:

$$\begin{pmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{pmatrix} \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{1.34}$$

We can express the following relations:

$$\xi_x = \frac{\begin{vmatrix} 1 & x_\eta \\ 0 & y_\eta \end{vmatrix}}{\begin{vmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{vmatrix}} \qquad \Rightarrow J\xi_x = y_\eta \quad J\xi_y = -x_\eta \quad J\eta_x = -y_\xi \quad J\eta_y = x_\xi \tag{1.35}$$

$$\Rightarrow \frac{\partial (J\xi_x)}{\partial \xi} + \frac{\partial (J\eta_x)}{\partial \eta} = 0 \qquad \frac{\partial (J\xi_y)}{\partial \xi} + \frac{\partial (J\eta_y)}{\partial \eta} = 0$$

We can multiply by J (1.33) and find:

$$\begin{aligned} J\frac{\partial F_x}{\partial x} &= J\xi_x\frac{\partial F_x}{\partial \xi} + J\eta_x\frac{\partial F_x}{\partial \eta} = \frac{\partial (J\xi_x F_x)}{\partial \xi} + \frac{\partial (J\eta_x F_x)}{\partial \eta} - F_x\underbrace{\left(\frac{\partial (J\xi_x)}{\partial \xi} + \frac{\partial (J\eta_x)}{\partial \eta}\right)}_{=0} \\ J\frac{\partial F_y}{\partial y} &= J\xi_y\frac{\partial F_y}{\partial \xi} + J\eta_y\frac{\partial F_y}{\partial \eta} = \frac{\partial (J\xi_y F_y)}{\partial \xi} + \frac{\partial (J\eta_y F_y)}{\partial \eta} - F_y\underbrace{\left(\frac{\partial (J\xi_y)}{\partial \xi} + \frac{\partial (J\eta_y)}{\partial \eta}\right)}_{=0} \end{aligned} \tag{1.36}$$

So that by multiplying the transformed Euler equation (conservative) by J and replacing we get:

$$J\frac{\partial U}{\partial t} + \frac{\partial (J\xi_x F_x + J\xi_y F_y)}{\partial \xi} + \frac{\partial (J\eta_x F_x + J\eta_y F_y)}{\partial \eta} = 0 \tag{1.37}$$

We can now discretize this equation using centered finite differences:

$$\begin{aligned} J\frac{\partial U_{ij}}{\partial t} &+ \frac{(J\xi_x F_x + J\xi_y F_y)_{i+\frac{1}{2}\,j} - (J\xi_x F_x + J\xi_y F_y)_{i-\frac{1}{2}\,j}}{\Delta \xi} \\ &+ \frac{(J\eta_x F_x + J\eta_y F_y)_{i\,j+\frac{1}{2}} - (J\eta_x F_x + J\eta_y F_y)_{i\,j-\frac{1}{2}}}{\Delta \eta} = 0 \end{aligned} \tag{1.38}$$

where $\Delta\xi = \Delta\eta = 1$. We can easily rewrite this by defining the middle point as average:

$$\begin{aligned} J\frac{\partial U_{ij}}{\partial t} &+ (J\nabla\xi)_{i+\frac{1}{2}\,j}\frac{\vec{F}_{ij} + \vec{F}_{i+1\,j}}{2} - (J\nabla\xi)_{i-\frac{1}{2}\,j}\frac{\vec{F}_{i-1\,j} + \vec{F}_{ij}}{2} \\ &+ (J\nabla\eta)_{i\,j+\frac{1}{2}}\frac{\vec{F}_{ij} + \vec{F}_{i\,j+1}}{2} - (J\nabla\eta)_{i\,j-\frac{1}{2}}\frac{\vec{F}_{i\,j-1} + \vec{F}_{ij}}{2} = 0 \end{aligned} \tag{1.39}$$

We can express the derivatives of the metric terms easily, for example for:

$$\begin{aligned} (J\xi_x)_{i+\frac{1}{2}\,j} &= (y_\eta)_{i+\frac{1}{2}\,j} = \frac{y_B - y_A}{\Delta\eta} = \Delta y_{AB} \\ (J\xi_y)_{i+\frac{1}{2}\,j} &= (-x_\eta)_{i+\frac{1}{2}\,j} = \frac{-x_B + x_A}{\Delta\eta} = -\Delta x_{AB} \end{aligned} \tag{1.40}$$

We found out that $(J\nabla\xi)_{i+\frac{1}{2}\,j} = (\vec{n}l)_{AB}$ and we can do the same for the others. To conclude, we have to proof that $J = \Omega_{ij}$ but this is obvious by computing the determinant and we know the expression of $\vec{x}_\xi$ and $\vec{x}_\eta$ at point $i \pm \frac{1}{2}\,j$ and $i\,j \pm \frac{1}{2}$ so we just have to make the average to have ij:

$$J_{ij} = |\vec{x}_\xi \times \vec{x}_\eta|_{ij} = \frac{1}{2}(\Delta\vec{x}_{DA} - \Delta\vec{x}_{BC} \times \Delta\vec{x}_{AB} - \Delta\vec{x}_{CD}) \tag{1.41}$$

We can conclude that FV is the generalization of FD in changed coordinate. The advantage of the finite volume is that we can now deal with any quadrilateral.

1.3 Finite element method

It is based on a functional representation, a linear combination of basis or **shape functions** $u*(x) = \sum_{i=1}^{n} a_i v_i(x)$ and the parameters are the **coefficients** of the basis functions, that are determine such that we get the best approximation of the exact solution. It differs from Galerkin and Ritz method by the selection of a **piecewise polynomial interpolation function** as

shape functions. It is a method to construct generalized finite difference formulas. For example, the numerical solution parameters are values of the solution at particular points (nodes) and the discretized equations have a local character as in FD.

The convergence of the method is primordial, the numerical solution must tend to the exact solution when the number of parameters tends to infinity.

1.3.1 Various form of a differential equation

We already saw that the conservation laws can be put under an integral form. But other forms exist, for example the **weak form** and the **variational form**.

The **differential form**, also called strong form (needs to be satisfied on all points of the domain and thus the derivatives too). Then we have the **integral form** developed by physicians and that does not require to be differentiable everywhere and thus allows the existence of discontinuities: **weak solutions**.

To illustrate the weak form, consider the elastic deformation of a bar fixed at his upper end and submitted to its own weight and a traction force at the other end. Taking x-axis pointing downward, the strong form is:

$$\frac{d}{dx}\left(k\frac{du}{dx}\right)+\mu g=0 \qquad u(0)=0 \quad k\frac{du}{dx}(L)-F=0 \tag{1.42}$$

where u is the displacement, k the rigidity (can be discontinuous) and μ the mass per unit length. As the equation must be satisfied everywhere, we can integrate the equation after having multiplied by a test function and add a null term due to the boundary condition:

$$-\int_0^L \nu(x)\left[\frac{d}{dx}\left(k\frac{du}{dx}\right)+\mu g\right]\,dx+\nu(L)\left[k\frac{du}{dx}(L)-F\right]=0 \tag{1.43}$$

And by integrating by parts $f=\nu(x)$ and $g'=\frac{d}{dx}(k\frac{du}{dx})$:

$$\int_0^L \left[\frac{d\nu}{dx}k\frac{du}{dx}-\mu g\nu\right]\,dx-\underbrace{\nu(0)k\frac{du}{dx}(0)}_{=0}-\nu(L)F=0 \tag{1.44}$$

where we make the underbraced term = 0 by choosing $\nu(0)=0$ in order to satisfy the boundary condition at the fixation. This is the weak form of the equation and is also found by application of the virtual work theorem. The highest differentiation order is here reduced from two to one and allows the first order derivative to be discontinuous. This example does not contain that unless the material property is discontinuous.

In fluid mechanics, shocks are discontinuities and consider the following equation in conservation form of a quasi-one dimensional nozzle flow:

$$\begin{gathered}\frac{d}{dx}\left(\frac{u^2}{2}\right)-xu=0 \qquad u(-1)=1; \quad u(1)=-0.5\\ u\frac{du}{dx}=xu\Leftrightarrow u=\frac{x^2}{2}+c\end{gathered} \tag{1.45}$$

By considering the two boundary conditions the discontinuity appears:

$$U_L = \frac{x^2+1}{L} \qquad U_R = -1 + \frac{x^2}{2} \tag{1.46}$$

The weak form of the previous equation is again obtained using the weighting function $\nu(x)$ that must now vanish at 1 and -1:

$$\int_{-1}^{1} \nu(x) \left[\frac{d}{dx}\left(\frac{u^2}{2}\right) - xu \right] dx = \int_{-1}^{1} \left[\frac{d\nu}{dx}\frac{u^2}{2} - ux \right] dx = 0a \tag{1.47}$$

where there is no longer the derivative of u. Come back to the bar, an extremum case can be to choose a variation for $\nu(x) = \delta u(x)$ such that:

$$\int_0^L \left[\frac{d\delta u}{dx} k \frac{du}{dx} - \mu g \delta u \right] dx - \delta u(L) F = 0 = \delta \left[\int_0^L \left[\frac{k}{2}\left(\frac{du}{dx}\right)^2 - \mu g u \right] dx - u(L)F \right] \tag{1.48}$$

This can be interpreted as the variation of the total energy and the solution is the one that gives stationary energy.

1.3.2 Shape functions - finite element interpolation

As explained, shape functions are piecewise polynomial interpolation functions and its parameters are values of the numerical solution in certain points. Here are the principles to construct these shape functions:

- The domain is divided into a set of non-overlapping simple polygons or polyhedra (the edges are not forced to be straight);
- To each domain we associate some points called **nodes** and we get an **element**. In almost all cases, the nodes include the vertices but could also be on the edges or inside the element.;
- To each node N is associated a function $\nu_n(x)$ which is defined locally within the element as a polynomial interpolation function. It must vanish at other nodes:

$$\forall P \in \Omega_e : \nu_n(xp) = \delta_{np} \tag{1.49}$$

This definition of the interpolation function has several consequences:

— The order of the polynomial is directly linked to the number of nodes. For example for a triangular 3 nodes element we will use a linear interpolation in 2 space variables (P1), while for quadratic polynomials 6 nodes are required since we have 6 coefficients in the interpolation (P2).

— The shape function is defined locally for each element. The global basis function associated to a node is thus simply the function equal to the local basis function on each element. If the node belongs to one element the global basis function is the local basis function of the element, if it belongs to several, the global basis function is made of all the local shape function of all the elements it belongs to. A consequence is that the global function associated to node N vanish on all other nodes (called compact support).

Also since the basis functions are uniquely defined on each element and since they coincide on a mesh belonging to several element, these are continuous functions.

— The coefficients of the interpolations are values of the numerical solution at the corresponding node:

$$u^*(x_i) = \sum_{j=1}^{n} a_j \nu_j(x_i) = a_i \tag{1.50}$$

We shall call them $a_i = u_i$.

We are going to consider some example in 1D, where we have to divide the the domain into set of intervals and associate nodes. The attribution of nodes differs from polynomial order.

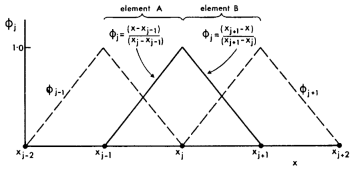

Figure 1.8

P1 element First order polynomial has two coefficients $(a + bx)$ thus we need two nodes per element to determine them all, we take them on vertices. If we call the local shape functions on element A and B, the global shape function around node j is made of the two.

$$\phi_i^A = \frac{x - x_{j-1}}{x_j - x_{j-1}} \qquad \phi_i^B = \frac{x_{j+1} - x}{x_{j+1} - x_j} \tag{1.51}$$

P2 element We have now a 2nd order polynomial $(a + bx + cx^2)$ with 3 coefficients and thus 3 nodes are needed. In general the third node is placed at the middle of the element. The shape functions can be simply chosen via the Lagrange interpolation, for example:

$$\begin{gathered}\phi_j^A = \frac{(x_{j+2}-x)(x_{j+1}-x)}{(x_{j+2}-x_j)(x_{j+1}-x_j)} \qquad \phi_{j+1}^A = \frac{(x_{j+2}-x)(x-x_j)}{(x_{j+2}-x_{j+1})(x_{j+1}-x_j)} \\ \phi_{j+2}^A = \frac{(x_{j+1}-x)(x-x_j)}{(x_{j+1}-x_{j+2})(x_{j+2}-x_j)}\end{gathered} \tag{1.52}$$

or for higher dimensions, using a parent element. The parent element is defined in ξ, η for the higher dimension case and in ξ coordinate for 1D where for P2 element the nodes are placed at 0, 1/2 and 1. The shape functions in the normalized coordinates are:

$$\phi_1 = (2\xi - 1)(\xi - 1) \qquad \phi_2 = 4\xi(1 - \xi) \qquad \phi_3 = \xi(2\xi - 1) \tag{1.53}$$

The same shape functions can be used for the transformation, in higher dimension we have:

$$x(\xi, \eta) = \sum x_i \phi_i(x, y) \qquad y(\xi, \eta) = \sum y_i \phi_i(x, y) \tag{1.54}$$

We speak about **isoparametric mapping** in this case since the polynomial order is equal to the number of nodes. If lower order polynomials are used we speak of subparametric mapping.

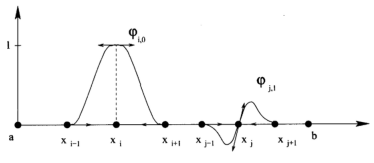

Figure 1.9

P3 Hermite element We have $a + bx + cx^2 + dx^3$, we need thus 4 nodes per element. We could just do as before so use Lagrange interpolation with the 2 vertices and two internal nodes, but it is also possible to use Hermite element: we only use the vertices and the numerical parameters are not only solution at nodes but also of its derivative. We have 2 dof at each node so that the shape functions in the normalized element are:

$$\phi_{0,0} = (2\xi + 1)(\xi - 1) \quad \phi_{0,1} = \xi(\xi - 1)^2 \quad \phi_{1,0} = (3 - 2\xi)\xi^2 \quad \phi_{1,1} = (\xi - 1)\xi^2 \tag{1.55}$$

In P1 and P2 the basis functions have discontinuous derivatives. The advantage of Hermite is to ensure also the continuity of the derivatives at the nodes. The other advantage is counting

the number of dof. Suppose an interval from 0 to L and n elements in the interval, how many unknowns with Lagrange? $\#dof = n + 1 + 2internals = n + 1 + 2n = 3n + 1 = \mathcal{O}(3n)$ while it is $\mathcal{O}(2n)$ for Hermite. We have the same accuracy at a lower cost.

1.3.3 Discretization - extremum form: Ritz method

In some cases we do have a variational form, we can find the numerical solution by imposing it to make the variation stationary within the set of numerical solutions of the chosen form:

$$\frac{\partial E(u^h)}{\partial u_i} = 0 \qquad i = 1, \ldots, n \tag{1.56}$$

We have thus n equations and n unknowns and the solution will be the best in the energy sense. Generally we have something of the form:

$$\begin{aligned} E(u) &= \int_\Omega F(u, \nabla u)\, d\Omega + \int_{\partial\Omega} g(u)\, d\Gamma \\ \Rightarrow \frac{\partial E(u^h)}{\partial u_i} &= \int_\Omega \left[\frac{\partial F}{\partial u}\frac{\partial u^h}{\partial u_i} + \frac{\partial F}{\partial u_{x_p}}\frac{\partial}{\partial u_i}\left(\frac{\partial u^h}{\partial x_p}\right)\right] d\Omega + \int_{\partial\Omega} \frac{\partial g(u)}{\partial u}\frac{\partial u^h}{\partial u_i}\, d\Gamma = 0 \\ &= \int_\Omega \left[\frac{\partial F}{\partial u} v_i + \frac{\partial F}{\partial u_{x_p}}\frac{\partial v_i}{\partial x_p}\right] d\Omega + \int_{\partial\Omega} \frac{\partial g(u)}{\partial u} v_i\, d\Gamma = 0 \end{aligned} \tag{1.57}$$

The last line of the equation is obtained by remembering $u^h(x) = \sum_{i=1}^n u_i v_i(x)$. Remark that the last line is the weak form of the equation with v_i as test function.

1.3.4 Discretization - weak form: weighted residual method

All the differential equations cannot be put under variational form and the previous case is not always applicable. In contrast weak form exists in all case and can be used as basis for discretization. Symbolically if we have $D(u) = 0$, for a numerical solution u^h the equation will not vanish. We will have a residual $D(u^h) = r^h$. It is thus logical to determine the coefficients u_i so that the residual is minimized. The **least-square** definition consists in minimizing the residual quadratic nom:

$$J(u^h) = \int_\Omega (r^h)^2\, d\Omega \tag{1.58}$$

where for simplicity we considered prescribed solution at the boundary but in general case the boundary condition terms should be added. The minimization consists in:

$$\frac{\partial J(u^h)}{\partial u_i} = \int_\Omega 2\frac{\partial r^h}{\partial u_i} r^h\, d\Omega = 0 \qquad i = 1, \ldots, n \tag{1.59}$$

the weighting functions are $w_i^{LS} = 2(\partial r^h/\partial u_i)$, and after integration we find the weak form of the differential equation:

$$\int_\Omega [vF(u, \nabla u) + \nabla v.\vec{g}(u, \nabla u)]\, d\Omega + \int_{\Gamma=\partial\Omega} vH(u)\, d\Gamma = 0 \tag{1.60}$$

were we choose $v_i = w_i(x)$. You take as many weighting functions as parameters. Let's enumerate a certain number of requirements the w_i should satisfy:

- They should be in the same number of the numerical parameters/shape functions to provide closed algebraic system;

- They should form a complete set: if you have a domain Ω the weighting functions should fill the whole domain, no empty space.

Discretization is not unique it depends on the choice of weight.

Galerkin method

The shape functions satisfy the two criteria. If we do that we refind the Ritz method. When a problem can be cast in extremum form, Galerkin method is similar to the Ritz method and is optimal in energy sense. Since the weighting functions belongs to the same functional space than the numerical solution, it requires continuity of the shape functions only up to one order less than that of the highest derivative in the weak form.

Least squares method

Has been seen at the beginning of the section. It has the advantage to be tackled by classical minimization methods and to solve the instability problems of the Galerkin method for convection problems. But in contrast it requires C^1 continuities for second order derivatives, costly.

Point collocation method

One of the drawbacks of previous methods is to perform computation of comlpex integrals. This can be avoided by using Dirac distributions as weighting functions $w_i(x) = \delta(x - x_i)$ so that the discrete equations become $r^h(x) = 0$. The main disadvantages are that to require a high order approximation and to be less accurate.

Subdomain collocation method

This is a finite volume-like method whereas the previous is a finite element-like method. We choose piecewise constant functions equal to one on a subdomain Ω_i and zero elsewhere:

$$\int_{\Omega_i} r^h \, d\Omega = 0 \tag{1.61}$$

This approach offers advantage when the equation to be solved is a conservation law so that the integral can be transformed into surface integral through Gauss theorem. This is known as control-volume based finite element method. The number of subdomains must be equal to the number of nodes. The method is similar to cell vertex centered or vertex centered method since the subdomain is chosen as the set of nodes like them.

Petrov-Galerkin method

When it is not one of the previous method and we use several weighting we speak of that.

1.4 Spectral methods

1.4.1 Representation

Just like finite element, it is based on a functional representation of the solution $u^*(x) = \sum_{i=1}^{n} a_i v_i(x)$. The difference is that instead of choosing a piecewise polynomial as shape function, we choose trigonometric functions or families of orthogonal functions. For example in lifting line theory it was logical to express $\Gamma(\theta) = \sum m = 1^N a_m \sin mx$ with $\Gamma(0) = \Gamma(\pi) = 0$, use

of truncated Fourrier serie. As long as the solution is smooth it provides accurate values for few terms in the serie. There are 2 fundamental differences, basic functions are infinitely differentiable, we don't have to worry about the continuity. In finite elements we had C^0 continuity. The convergence is much faster for spectral methods when number of elements increases: FE: $\epsilon \propto h^{k+1}$, spectral: $\epsilon \propto \exp(-1/h)$. The approximation can be noted as:

$$u^*(x) = \sum_{k=-\frac{n}{2}+1}^{\frac{n}{2}} \hat{u}_k \exp\left(\frac{2\pi I k x}{L}\right) \tag{1.62}$$

For periodic problems, v_i are trigonometric functions. For non periodic problems we use families of orthogonal polynomials such as Chebyskev, Legendre, Laguerre. Spectral methods is very limited to some problems, it is used for numerical computations around the atmosphere (forecasting).

But there is a price to pay for that, the shape functions are not defined locally but globally and the shape functions do not represent the solution value in some certain point. They are non zero over the whole domain, so the system of equations is not sparse, all the parameters are coupled. Basis functions are not of compact support!

Discretization

Same as FE. Since the basis functions are infinitely differentiable we don't have problems with collocations.

Chapter 2

Elements of PDE's

Convergence was an interesting issue but the type of the equation we have to solve is also important. This is why we study the elements of the theory of PDE's

2.1 Quasi-linear equations – Conservative form

A quasi-linear equation is an equation that is linear in the highest derivatives. We start with a first order equation.

General form of a first order quasi-linear equation in two variables

We look for a function of $u(x, y)$ involving only first order derivatives in 2 space coordinates. And typically we have a linear combination of x and y derivatives and the coefficients may depend on u too:

$$P(x, y, u)\frac{\partial u}{\partial x} + Q(x, y, u)\frac{\partial u}{\partial y} = R(x, y, u) \tag{2.1}$$

This doesn't mean that the equation is linear, for example $\frac{\partial u}{\partial x} + u\frac{\partial u}{\partial y} = S(x, y, u)$: non linear Burger's equation.

General form of a second order quasi-linear equation in two variables

The second order equation in 2 space variables is:

$$\begin{aligned} P\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)\frac{\partial u^2}{\partial x^2} + 2S\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)\frac{\partial u^2}{\partial x \partial y} \\ + T\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)\frac{\partial u^2}{\partial y^2} = H\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) \end{aligned} \tag{2.2}$$

In most applications in fluid mechanics, the factors multiplying the higher order derivatives do not depend explicitly on the independent variables x, y. But they will depend implicitly since u is a function of x, y. Let's give an example of this, 2D potential equation for compressible flows:

$$(a^2 - u^2)\frac{\partial^2 \varphi}{\partial x^2} - 2uv\frac{\partial^2 \varphi}{\partial x \partial y} + (a^2 - v^2)\frac{\partial^2 \varphi}{\partial y^2} = 0 \tag{2.3}$$

Quasi-linear equations can appears under the conservative/divergence form:

$$\frac{\partial g_x}{\partial x} + \frac{\partial g_y}{\partial y} (= \nabla . \vec{g}) = S(x, y, u) \tag{2.4}$$

where $g_x, g_y = \vec{g}(x, y, u) = \vec{g}\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$ for respectively a first order and a second order equation. Notice that it is possible to recover from here the quasi-linear form. Let's call $\tilde{g}_x(x, y) = \hat{g}_x(x, y, u(x, y))$, first order equation, then we have that by chain rule:

$$\frac{\partial \tilde{g}_x}{\partial x} = \frac{\partial \hat{g}_x}{\partial x} + \frac{\partial \hat{g}_x}{\partial u}\frac{\partial u}{\partial x} \qquad \frac{\partial \tilde{g}_y}{\partial y} = \frac{\partial \hat{g}_y}{\partial y} + \frac{\partial \hat{g}_y}{\partial u}\frac{\partial u}{\partial y} \tag{2.5}$$

Then the sum of the two gives:

$$\begin{aligned} \frac{\partial \tilde{g}_x}{\partial x} + \frac{\partial \tilde{g}_y}{\partial y} &= \frac{\partial \hat{g}_x}{\partial x} + \frac{\partial \hat{g}_y}{\partial y} + \frac{\partial \hat{g}_x}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial \hat{g}_y}{\partial u}\frac{\partial u}{\partial y} \\ \frac{\partial \hat{g}_x}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial \hat{g}_y}{\partial u}\frac{\partial u}{\partial y} &= S - \frac{\partial \tilde{g}_x}{\partial x} - \frac{\partial \tilde{g}_y}{\partial y} \end{aligned} \tag{2.6}$$

where we refined our coefficients P and Q. For a first order equation in 2 space variables the converse is true as well. Indeed, defining $P = \frac{\partial \hat{P}}{\partial u}$ and $Q = \frac{\partial \hat{Q}}{\partial u}$, the chain rule is:

$$\frac{\partial \tilde{P}}{\partial x} = \frac{\partial \hat{P}}{\partial x} + \frac{\partial \hat{P}}{\partial u}\frac{\partial u}{\partial x} = \frac{\partial \hat{P}}{\partial x} + P\frac{\partial u}{\partial x} \qquad \Rightarrow P\frac{\partial u}{\partial x} = \frac{\partial \tilde{P}}{\partial x} - \frac{\partial \hat{P}}{\partial x} \tag{2.7}$$

Replacing these in the general first order equation form (2.1), we find back the conservative form:

$$\frac{\partial \tilde{P}}{\partial x} + \frac{\partial \tilde{Q}}{\partial y} = S + \frac{\partial \hat{P}}{\partial x} + \frac{\partial \hat{Q}}{\partial y} \tag{2.8}$$

For a second order equation, this is not always possible.

2.2 Characteristic surfaces and wave-like solutions

2.2.1 First order scalar equation in m independent variables

It means that we have a linear combination such as:

$$a_i \frac{\partial u}{\partial x_i} = 0 \tag{2.9}$$

For simplicity we take the source term = 0. We suppose to solve an initial value problem (Cauchy problem). We imagine that the solution is known on some hyper-surface S^* (in 2D a curve) of equation $F(x_i) = 0$. Does this problem have one and only one solution?

The value of u on the surface is called the **trace** of the solution on the hyper surface and it is specified. Imagine that we can construct a function $\varphi(x, y)$ such that it is equal to u on the surface. Typically if we think in 2D, we specify u on a curve and we elongate it arbitrarily. Because the function $\varphi = u$ on the surface, then the tangential derivative of φ and u are the same. If we construct the function $\varphi - \lambda F$ is also equal to u on the surface whatever the value of λ. In other words, I know u on the curve, I construct a function φ and then I say that all the function $\varphi - \lambda F$ are the same as u, so we have an infinite number of solution on the surface. But there exists one λ for which the normal derivative will be the same as the normal derivative of u:

$$\nabla\varphi - \lambda\nabla F = \nabla u \qquad \Leftrightarrow \frac{\partial\varphi}{\partial x_i} - \lambda\frac{\partial F}{\partial x_i} = \frac{\partial u}{\partial x_i} \quad i = 1, \ldots, m \tag{2.10}$$

The unknowns in this equation are the partial derivatives of u but are given by $a_i\frac{\partial u}{\partial x_i} = 0$. We have thus a system of $m+1$ equations and $m+1$ unknowns. The system can be put under matrix form as:

$$\begin{pmatrix} 1 & \ldots & & & \frac{\partial F}{\partial x_1} \\ & 1 & & & \frac{\partial F}{\partial x_2} \\ \vdots & & & & \vdots \\ a_1 & a_2 & \ldots & a_m & 0 \end{pmatrix} \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \vdots \\ \frac{\partial u}{\partial x_m} \\ \lambda \end{pmatrix} = \begin{pmatrix} \frac{\partial\varphi}{\partial x_1} \\ \vdots \\ \frac{\partial\varphi}{\partial x_m} \\ 0 \end{pmatrix} \tag{2.11}$$

The system has one and only one solution unless if the determinant is equal to 0, unless the surface S is such that $a_i\frac{\partial F}{\partial x_i} = 0 = \vec{a}.\nabla F$. In fact ∇F is parallel to the normal to the surface because all the tangential derivatives are 0 and the only component that cannot be 0 is the normal derivative, so $\nabla F \propto \vec{n}$. And if we say $\vec{a}.\vec{n} = 0$, this means that the surface is tangent to $\vec{a}$. These $\vec{a}$ are called **characteristic lines** of a **characteristic surface**. The response to the question is thus that yes the solution is unique unless if the surface is a characteristic surface for which case we have no solution or an infinity of solutions if compatible. In 2D this is a characteristic curve.

The original equation (2.9) admits solutions of the form: $u = \hat{u}\exp(IF(x_i))$ where $F(x_i) = 0$ are equations of characteristic surfaces. Because if we compute $\frac{\partial u}{\partial x_i} = Iu\exp(IF(x_i))\frac{\partial F}{\partial x_i} = Iu\frac{\partial F}{\partial x_i}$ and thus

$$a_i\frac{\partial u}{\partial x_i} = IUa_i\frac{\partial F}{\partial x_i} = 0 \tag{2.12}$$

which is identically satisfied if $\vec{a}.\vec{n} = 0$. Lines of constant F are wave fronts of wave-like solutions since we have $\exp(IF(x_i))$ similar to $\exp(kx - \omega t)$. A special case is when $a_i = cst \rightarrow \frac{\partial F}{\partial x_i = n_i}$ where we have planar wave that propagates without dilatation or damping ($\hat{u} = cst$).

2.2.2 Second order equations in one unknown in two dimension

Consider the second order equation in two variables:

$$R\frac{\partial^2 u}{\partial x^2} + 2S\frac{\partial^2 u}{\partial x\partial y} + T\frac{\partial^2 u}{\partial y^2} = 0 \tag{2.13}$$

In this case we have to provide u and ∇u but if we know the surface we can compute ∇u. Let's call $p = \frac{\partial u}{\partial x}$ and $q = \frac{\partial u}{\partial y}$ on the curve. Let's call φ a function equal to p on C. It results that $\varphi = p + \lambda F$ on C and therefore there exists a value of λ such that $\nabla\varphi = \nabla p + \lambda\nabla F$. So we have also for the second variable:

$$\psi = q + \mu F \qquad \Rightarrow \nabla\psi = \nabla q + \mu\nabla F \tag{2.14}$$

The unknowns are λ, μ, the components of the gradients p and q. But we know that

$$\nabla p = \frac{\partial^2 u}{\partial x^2}\vec{e}_x + \frac{\partial^2 u}{\partial x\partial y}\vec{e}_y \qquad \nabla q = \frac{\partial^2 u}{\partial y^2}\vec{e}_y + \frac{\partial^2 u}{\partial x\partial y}\vec{e}_x \tag{2.15}$$

Again we have a system of equation

$$\begin{pmatrix} 1 & 0 & 0 & \frac{\partial F}{\partial x} & 0 \\ 0 & 1 & 0 & \frac{\partial F}{\partial y} & 0 \\ 0 & 1 & 0 & 0 & \frac{\partial F}{\partial x} \\ 0 & 0 & 1 & 0 & \frac{\partial F}{\partial y} \\ R & 2S & T & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{\partial p}{\partial x} \\ \frac{\partial p}{\partial y} = \frac{\partial q}{\partial x} \\ \frac{\partial q}{\partial y} \\ \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} \frac{\partial \varphi}{\partial x} \\ \frac{\partial \varphi}{\partial y} \\ \frac{\partial \psi}{\partial x} \\ 0 \\ 0 \end{pmatrix} \tag{2.16}$$

If we try to compute the determinant we will get

$$\det = R\frac{\partial F}{\partial x}\left(-\frac{\partial F}{\partial x}\right) - 2S\frac{\partial F}{\partial x}\frac{\partial F}{\partial y} + T\left(-\frac{\partial F}{\partial y}\right)\frac{\partial F}{\partial y} = -\left(\frac{\partial F}{\partial y}\right)^2\left[Rz^2 + 2Sz + T\right] \tag{2.17}$$

where $Z = \frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}} = \frac{n_x}{n_y}$. The line $F(x, y) = 0$ is a characteristic line when $Rz^2 + 2Sz + T = 0$. If $S^2 - RT > 0$ so we have 2 roots and thus 2 characteristic directions at each point. In the case $S^2 - RT < 0$ we have no real root no characteristic line for this equation. And if $S^2 - RT = 0$ we have 2 identical roots so 1 characteristic direction. If now we have a quadratic term like

$$Rx^2 + 2Sxy + Tx^2 = 0 \tag{2.18}$$

The first case would give a hyperbole, the second an elliptic equation and the last we have a parabolic equation.

Applicaition: potential flow equation

$$\begin{gathered} R = a^2 - u^2 \qquad S = -uv \qquad T = a^2 - v^2 \\ \Rightarrow S^2 - RT = u^2v^2 - (a^2 - u^2)(a^2 - v^2) = a^2[u^2 + v^2 - a^2] = a^4\left[M^2 - 1\right] \end{gathered} \tag{2.19}$$

We can see that when $M > 1$ hyperbolic, $M = 1$ parabolic, $M < 1$ elliptic, the parameters R, S, T change size within the domain.

2.2.3 System of first order equations in two dimensions

We can write it in the form of a system of n equations and n unknowns:

$$A_x\frac{\partial U}{\partial x} + A_y\frac{\partial U}{\partial y} = 0 \tag{2.20}$$

It is now U that is provided on the curve C of equation $F(x, y) = 0$. We call V a vector function which is identical to U on C. We are going to have a vector of Lagrange multipliers Θ:

$$V = U + \Theta F \text{ on C} \qquad \Rightarrow \frac{\partial V}{\partial x} = \frac{\partial U}{\partial x} + \Theta\frac{\partial F}{\partial x} \qquad \frac{\partial V}{\partial y} = \frac{\partial U}{\partial y} + \Theta\frac{\partial F}{\partial y} \tag{2.21}$$

We will again have the matrices:

$$\begin{pmatrix} I & 0 & \frac{\partial F}{\partial x}I \\ 0 & I & \frac{\partial F}{\partial y}I \\ A_x & A_y & 0 \end{pmatrix} \begin{pmatrix} \frac{\partial U}{\partial x} \\ \frac{\partial U}{\partial y} \\ \Theta \end{pmatrix} = \begin{pmatrix} \frac{\partial V}{\partial x} \\ \frac{\partial V}{\partial y} \\ 0 \end{pmatrix} \tag{2.22}$$

Again we have to compute the determinant. The characteristic directions are $|A_xn_x + A_yn_y| = 0$ and if we factorize:

$$n_x|A_x + \tilde{\lambda}A_y| = 0 \qquad n_y\left|-A_x\frac{-n_x}{n_y} + A_y\right| = 0 \tag{2.23}$$

where $\frac{-n_x}{n_y}$ is the slope of the tangent to the characteristic lines. We end up with an eigenvalue problem

$$|A_y - \lambda A_x| = 0 \tag{2.24}$$

if $\det(A_x) \neq 0$ so $|A_y - \lambda A_x| = |A_x||A_x^{-1}A_y - \lambda I| = 0$. We have n real roots, n lin indep real eigenvectors (hyperbolic) or n real roots, m<n lin indep real eigenvectors (parabolic) [1]. We also have the case elliptic problem with n complex roots. Other possibilities (complex + real roots) = hybrid, as previous mixed problems.

Notice that if we if we denote Λ the diagonal matrix of eigenvalues and L the matrix of left (line) eigenvectors, we can write:

$$L(A_y - \lambda A_x) = 0 \Leftrightarrow LA_y - \Lambda LA_x = 0 \qquad \Rightarrow LA_x\frac{\partial U}{\partial x} + LA_y\frac{\partial U}{\partial y} = 0 \Leftrightarrow \frac{\partial W}{\partial x} + \Lambda\frac{\partial W}{\partial y} = 0 \tag{2.25}$$

In the case of second order equations we would have:

$$\begin{gathered} R\frac{\partial^2 u}{\partial x^2} + 2S\frac{\partial^2 u}{\partial x \partial y} + T\frac{\partial^2 u}{\partial y^2} = 0 \qquad \Leftrightarrow R\frac{\partial p}{\partial x} + S\left(\frac{\partial p}{\partial y} + \frac{\partial q}{\partial x}\right) + T\frac{\partial q}{\partial y} = 0 \\ \begin{bmatrix} R & S \\ 0 & 1 \end{bmatrix} \frac{\partial}{\partial x}\begin{pmatrix} p \\ q \end{pmatrix} + \begin{pmatrix} S & T \\ 1 & 0 \end{pmatrix}\frac{\partial q}{\partial x} - \frac{\partial p}{\partial y} = 0 \end{gathered} \tag{2.26}$$

2.2.4 Systems of n equations in m independent variables

In that case the generalization is:

$$A_i\frac{\partial U}{\partial x_i} = 0 \tag{2.27}$$

where A is n by n matrix and U a n by 1 vector. The equations for the characteristic surfaces are now:

$$|A_i n_i| = 0 \tag{2.28}$$

We have again an eigenvalue problem, but it differs from previous ones since among the m components of $\vec{n}$, $m-1$ can be chosen arbitrarily, defining the nature of the problem, relatively to the variable associated to the component to determine. It will be respectively hyperbolic, parabolic, elliptic or hybrid if it is so for all the directions in the subspace of dimension $m-1$.

In general, if the problem is hyperbolic wrt one variable, it is indeterminate wrt the other variables, this is due to the fact that the component to be determined takes value in an interval wrt the first variable. As the problem admits wavelike solutions it can be named hyperbolic. Since for unsteady problems time often plays the role of privileged variable, it is common to designate the variable for which the problem is hyperbolic as **time-like** variable. In contrast, a problem elliptic wrt one variable is also for the others.

1. This is what happens when all derivatives of one component of U with respect to one independent variable x or y are missing.

EXAMPLE: To illustrate suppose that we have:

$$|A_x n_x + A_y n_y + A_z n_z| = 0 \tag{2.29}$$

We can write this in the following way:

$$\left|\frac{A_x n_x + A_y n_y}{n_z} + A_z\right| = 0 \qquad \Rightarrow \left|\frac{A_x n_x + A_y n_y}{\sqrt{n_x^2 + n_y^2}} - \frac{\sqrt{n_x^2 + n_y^2}}{n_z} + A_z\right| = 0 \tag{2.30}$$

where in fact the first term is the direction of n and the second term plays the role of the previous λ. If it is in 4D, we would have 2 angular directions defining the n direction. $m-2$ variables can be chosen arbitrarily and the last is given by the eigenvalue problem. The system is hyperbolic wrt to A_z if the eigenvalues are all real and we have a complete set of eigenvectors for all directions whatever the direction. If in contrast they are all complex whatever the direction, it is indetermined. The Euler equations will be:

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} = 0 \qquad U = \begin{pmatrix} \rho \\ \rho u \\ \rho E \end{pmatrix} \quad F = \begin{pmatrix} \rho u \\ p + \rho u 2 \\ \rho u H \end{pmatrix} \tag{2.31}$$

in 1D unsteady, z is an independent variable. We also know that $p = \rho RT = \rho(c_p - c_v)T$ and $z = c_v T = \rho(\gamma - 1)c_v T$ and $p = \rho(\gamma - 1)e, E = e + \frac{u^2}{2}, H = E + \frac{p}{\rho}$, so that we have:

$$\frac{\partial F}{\partial U} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{\partial p}{\partial e} - u^2 & 2u + \frac{\partial p}{\partial \rho u} & \frac{\partial p}{\partial \rho E} \\ \times & \times & \times \end{pmatrix} \tag{2.32}$$

And we see that the eigenvalues are real: $u + a, u, u - a$. In 2D the U and F vectors are 4 and thus the matrix will be 4×4 :

$$\begin{aligned} 2D: &\quad \frac{\partial U}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} = 0 \\ 3D: &\quad \frac{\partial U}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = 0 \end{aligned} \tag{2.33}$$

They are both unsteady with 3 and 4 independant variables. We see that these are hyperbolic wrt the time variable!

Conclusion

If the system has real eigenvalues and a complete set of real eigenvalues for all values of the arbitrary parameters (m-2) $\rightarrow$ the system is hyperbolic wrt the variable of interest $\rightarrow$ this variable plays a special role, it is the evolution or time-like variable.

If the system has only complex eigenvalues the equation is elliptic. In that case, it is generally elliptic wrt all variables.

2.2.5 Notion of well posed problem

Problem that has one and only one solution, depending continuously on the prescribed initial/boundary data. The last condition is primordial, if the initial conditions vary infinitesimally, the solution may only vary infinitesimally. The initial conditions choice will define if the

problem is well or ill-posed. The way to define them vary with the type of the problem. Let's illustrate with Laplace's equation:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{2.34}$$

and we compute the solution on the domain Ω which is the right half plane where $x \geq 0$. The boundary of the domain is $u(0, y) = 0$ and $\frac{\partial u}{\partial x}(0, y) = h(y)$. Since it is elliptic equation it has no real characteristic curve so we know that the curves are not characteristic and so that we have one and only one solution. If we take:

$$h(y) = 0 \Rightarrow u = 0 \qquad h(y) = \frac{\sin ny}{n} \Rightarrow u = \frac{1}{n^2} \sinh nx \sin ny \tag{2.35}$$

As $n \to \infty, h \to 0$, but for any non-zero x the solution will vary a lot. When we are never sure that the solution depends continuously on the boundary condition and when we have uncertainties on the boundary conditions we can never be sure of the solution. This is why if we have 2 boundary we have to impose one condition on each boundary.

2.3 Properties of hyperbolic equations

2.3.1 Nature of the solution - Riemann invariants

For simplicity, consider a system of n first order equations in 2 independent variables $|A_x n_x + A_y n_y| = 0 \Leftrightarrow |A_y - \lambda A_x| = 0$ where $\lambda = -n_x/n_y$. If A_x is positive definite regular matrix:

$$|A_x(A_x^{-1}A_y - \lambda I)| = 0 \qquad \Rightarrow |A_x||\underbrace{A_x^{-1}A_y}_{A} - \lambda I| = 0 \tag{2.36}$$

The λ_s are the eigenvalues of A. With $|A - \lambda I|v = 0 \Rightarrow Av = \lambda v$ where v are the right eigenvectors and l the line eigenvectors: $lA = \lambda l \Leftrightarrow (lA)^t = \lambda l^t$. We have the following algebraic identities:

$$Av_1 = \lambda_1 v_1 \quad Av_2 = \lambda_2 v_2 \ldots \qquad \Rightarrow AR = R\Lambda \Leftrightarrow R^{-1}AR = \Lambda \Leftrightarrow R^{-1}A = \Lambda R^{-1} \tag{2.37}$$

We find the relation $L = R^{-1}$. Let's imagine that the initial problem to solve is (as in aerodynamics course - Riemann):

$$A_x \frac{\partial U}{\partial x} + A_y \frac{\partial U}{\partial y} = S \qquad \Leftrightarrow \frac{\partial U}{\partial x} + A\frac{\partial U}{\partial y} = \underbrace{A_x^{-1}S}_{G} \Leftrightarrow L\frac{\partial U}{\partial x} + \Lambda L\frac{\partial U}{\partial y} = LG \tag{2.38}$$

If now we expand first line:

$$\begin{aligned} l_{11}\frac{\partial u_1}{\partial x} + l_{12}\frac{\partial u_2}{\partial x} + \cdots + l_{1n}\frac{\partial u_n}{\partial x} + \lambda_1\left[l_{11}\frac{\partial u_1}{\partial y} + \ldots\right] &= (LG)_1 \\ l_{11}\frac{\partial u_1}{\partial x} + \lambda_1 l_{11}\frac{\partial u_1}{\partial y} = l_{11}\left(\frac{\partial u_1}{\partial x} + \lambda_1\frac{\partial u_1}{\partial y}\right) &= l_{11}\left(\frac{\partial u_1}{\partial x} + \tan\theta\frac{\partial u_1}{\partial y}\right) \\ &= \frac{l_{11}}{\cos\theta}\left(\cos\theta\frac{\partial u_1}{\partial x} + \sin\theta\frac{\partial u_1}{\partial y}\right) \end{aligned} \tag{2.39}$$

where $\frac{1}{\cos\theta} = \sqrt{1+\lambda^2}, \vec{e}_\theta.\nabla u = \frac{du_1}{ds_1}$ (directional derivative in the characteristic direction with slope λ_i). This can be done for every indexes and we get:

$$\frac{1}{\cos\theta}\left(l_{11}\frac{du_1}{ds_1}+\cdots+l_{1n}\frac{du_n}{ds_1}\right)=h_1 \tag{2.40}$$

For a system of n equations, the system transforms into a system of n ordinary differential equations:

$$l_{ij}\frac{du_j}{ds_i}=h_i\cos\theta_i=\frac{h_i}{\sqrt{1+\lambda_i^2}} \tag{2.41}$$

This is a set of ordinary differential equtions along the characteristic curve i. Can we simplify further? If l_{ij} do not depend explicitly on the independent variables x, y (they depend only on u_j, true when A does not depend on x, y), there may exist a function $f(u_j)$ (integrating factor) such that

$$fl_{ij}\frac{du_j}{ds_i}=\frac{dR_i}{ds_i} \tag{2.42}$$

The left hand side is an exact differential of a certain function R_i. With the conditions:

$$fl_{ij}=\frac{\partial R_i}{\partial u_j}\Rightarrow\frac{\partial fl_{ij}}{\partial u_k}=\frac{\partial fl_{ik}}{\partial u_j}\qquad fl_{ik}=\frac{\partial R_i}{\partial u_k} \tag{2.43}$$

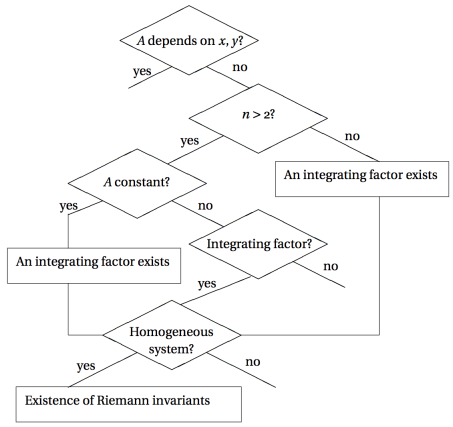

Figure 2.1

f is called an integrating factor. If $n\leq 2\rightarrow f$ always exists, if l_{ij} are constant $\forall n, R_i=l_{ij}u_j$ and if the system is homogeneous $g_i=0\rightarrow h_i=0\ \forall i$ we have:

$$\frac{dR_i}{ds_i}=\frac{fh_i}{\sqrt{1+\lambda_i^2}}=0\qquad\Rightarrow R_i=cst \tag{2.44}$$

R_i are the characteristic variables. The characteristic variables in previous section were the component of the vector $W=LA_yU$. When these are constant over a characteristic curve, they are called **Riemann invariants**. The summary is given on the figure.

Let's try to think about the properties of the solution. We restrict to the case of a system of two equations in two unknowns (so 2 characteristics). The solution is prescribed on the curve Γ, the solution at point P has to depend on all the solutions on the characteristic (integration) but these depend on the value of the solution in the section APB. In conclusion, the dark area is called the **region of dependence** of P. An analogous reasoning can denote the region behind P as **zone of influence of P**. The rest of the region is the **zone of silence**.

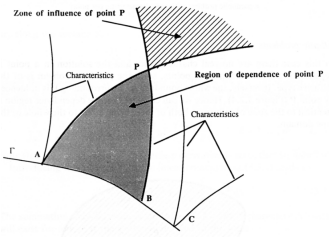

Figure 2.2

2.3.2 Well-posed problem for a hyperbolic system

The previous discussion allows to state that the Chauchy problem is well-posed for hyperbolic equations. Indeed, since the solution at point P only depends on the values of the solution upstream this point, we can compute it everywhere downstream of Γ. Furthermore, if the initial data are perturbed infinitesimally, the solution at any point will be perturbed so since wave-like solutions (property of hyperbolic equations) propagate without amplification or damping. As the solution is computed by progressing in the domain, the problem is an **evolution problem**.

Unsteady (time dependent) physical problems always lead to evolution problems. It is customary to note t the time-like variable with respect to which the problem is hyperbolic. Often the curve or surface over which the values of the dependent variables are specified is the surface $t = t_0$ and they are called initial conditions ($\rightarrow$ initial value problem).

Consider 1D inviscid compressible flow in a tube of length L. Continuity, x-momentum and energy equations tells that:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} = 0 \qquad \rho \frac{\partial u}{\partial t} + \rho u \frac{\partial u}{\partial x} = -\frac{\partial p}{\partial x} = -a^2 \frac{\partial \rho}{\partial x} \qquad \dot{s} = 0 \tag{2.45}$$

If initial data are homentropic (uniform initial entropy): $s = cst$. We can rewrite:

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} = 0 \qquad \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{a^2}{\rho} \frac{\partial \rho}{\partial x} = 0 \\ \Rightarrow &\frac{\partial}{\partial t} \begin{pmatrix} \rho \\ u \end{pmatrix} + \begin{pmatrix} u & \rho \\ \frac{a^2}{\rho} & u \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} \rho \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned} \tag{2.46}$$

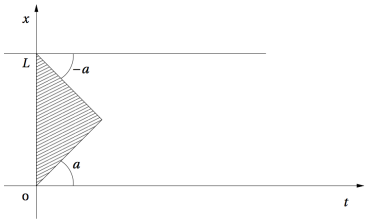

Figure 2.3

The eigenvalues of A are $\lambda = u \pm a$, confirming that it is hyperbolic. If we make the drawing, we have the graph of x in function of t and since the region of influence of the left boundary is limited to the triangle, we have to specify the upper and lower boundary conditions to be able to compute in all the tube. We have to supply as many information as the number of characteristic curves entering the domain, there is one on down boundary and one on top so we can prescribe the velocity for example.

2.3.3 Non-linear equations – Weak solutions

Consider the advection/reaction equation without source term:

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \tag{2.47}$$

Characteristic curves are lines tangent to the vector $\vec{a} = \vec{e}_t + u\vec{e}_x$, with slope u. Along the characteristics $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \sqrt{1+\lambda^2}\frac{du}{ds}$. The Riemann invariant is $u = cst$ on the curves. The equation of the characteristics are $x = x_0 + ut$ and the general solution of the homogeneous equation is $u = f(x - ut)$. The slope of the characteristics depend on the solution, it is possible that the lines cross. In that case on which curve do we have to rely to find the solution?

Consider the figure with the boundary conditions A, B, two cases can happen $A < B, A > B$. In the first case, the characteristics don't cross and an empty region appears, that can be filled by a **fan** $u = x/t$. In the second case, the characteristics cross. In that case the solution is not uniquely defined since it has to take both value A and B. The equation does not have a solution in the strong form. The weak form of the problem is:

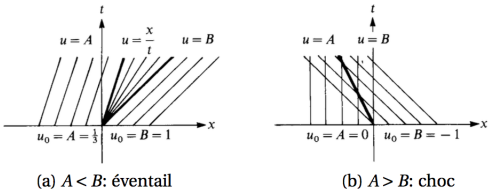

Figure 2.4

$$\int_0^\infty dt \int_{-\infty}^\infty \nu \left[\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right] dx = 0 \tag{2.48}$$

Then we can integrate by part and find that:

$$\int\int \nu \left[\frac{\partial u}{\partial t} + \frac{\partial u^2/2}{\partial x}\right] dx\, dt = \int\int \left[\frac{\partial uv}{\partial t} + \frac{\partial}{\partial x}\frac{u^2}{2}\nu\right] dx\, dt - \int\int \left[u\frac{\partial \nu}{\partial t} + \frac{u^2}{2}\frac{\partial \nu}{\partial x}\right] dx\, dt = 0$$
$$= -\int_0^\infty \int_{-\infty}^\infty \left[u\frac{\partial \nu}{\partial t} + \frac{u^2}{2}\frac{\partial \nu}{\partial x}\right] dx\, dt = -\int_{-\infty}^\infty u(0,x)\nu(0,x)\, dx \tag{2.49}$$

This form allows discontinuities, the question is now where are they located? For this let's consider the integral of the strong form:

$$\frac{d}{dt}\int_{x_1}^{x_2} u\, dx + \left[\frac{u^2}{2}\right]_{x_1}^{x_2} = 0 \tag{2.50}$$

and let's assume that there exists a discontinuity between x_1 and x_2 (jump from A to B). We can thus note x_s the location of the discontinuity and $\int_{x_1}^{x_2} = (x_s - x_1)A + (x_2 - x_s)B$ such that:

$$\frac{d}{dt}\int_{x_1}^{x_2} u\, dx = (A-B)\frac{dx_s}{dt} \equiv (A-B)\dot{x}_s \tag{2.51}$$

where $\dot{x}_s$ is the discontinuity displacement velocity. We can reconsider the equation noticing:

$$\dot{x}_s(B-A) = \left[\frac{u^2}{2}\right]_{x_1}^{x_2} = \frac{B^2 - A^2}{2} \qquad \Rightarrow \dot{x}_s = \frac{A+B}{2} \tag{2.52}$$

It is this solution which is represented on figure (b). This result is completely general for a general equation in the form:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{2.53}$$

The jump is called shock and the states on both side of the shock and the shock speed are linked by the jump condition also called Rankine-Hugoniot relation:

$$\dot{x}_s \underbrace{[u]}_{\text{jump of } u} = \underbrace{[f]}_{\text{jump of } f} \tag{2.54}$$

Let's come back to the case where $B > A$, the fan is a solution but with the weak form we can have the propagation of a discontinuity too now. One of the two solution is not valid in practice. For this we have to impose an additional condition:

For a shock to exist, the characteristics must lead into the shock when marching away from the Cauchy arc (here when t increases)

$$\left.\frac{dx}{dt}\right|_{\text{left char.}} > \dot{x}_s > \left.\frac{dx}{dt}\right|_{\text{right char.}} \Rightarrow \left.\frac{df(u)}{dt}\right|_{\text{left char.}} > \dot{x}_s > \left.\frac{df(u)}{dt}\right|_{\text{right char.}} \tag{2.55}$$

This condition is called **entropy condition** because in gas dynamics this is equivalent to the second principle of thermo. We have thus to get rid of the numerical solutions which violate the entropy condition.

Remark 1 It is important to use the correct conservative form. Indeed if we multiply Burger's equation by $2u$ we have:

$$\frac{\partial u^2}{\partial t} + \frac{\partial}{\partial x}\frac{2u^3}{3} \tag{2.56}$$

For this equation the jump condition reads $\dot{x}_s[u^2] = 2[u^3]/3$ which is a different shock speed, and thus leads to different results.

Remark 2 Consider the diffusive Burger's equation:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \alpha\frac{\partial^2 u}{\partial x^2} \tag{2.57}$$

Where the additional term is a diffusion term. It is possible to find analytical solutions for problem with boundary conditions like on **??**. The result when we let $\alpha \to 0$ is that when $u(0,x)$ is decreasing $(A > B)$, the solution tends towards a weak solution with shock whereas when it is increasing we tend towards a continuous solution. The solution satisfying the entropy condition is also is also the limit od the solution of the diffusive Burger's equation for $\alpha \to 0$. This is used: we introduce a diffusive term to eliminate the incorrect solutions (artificial viscosity or diffusivity techniques).

2.4 Properties of elliptical equations

2.4.1 Nature of the solution

The fundamental difference between elliptic and hyperbolic equations is that we have no real characteristics. In hyperbolic they have the property that the Fourrier modes propagates without damping or amplification. On the contrary, for elliptic equations amplified and damped Fourrier modes coexist. Let's see the Laplace's equation:

$$\Delta u = \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{2.58}$$

with the following boundary conditions $u(0,y) = g_1(y)$ and $\frac{\partial u}{\partial x}(0,y) = g_2(y)$. We will use Fourrier transforms in the y variable noted with a hat:

$$\begin{gathered} u(x,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{u}(x,\omega)e^{i\omega y}\,dy \qquad \frac{\partial u}{\partial y} = \frac{1}{2\pi}\int_{\infty}^{\infty} i\omega\hat{u}(x,\omega)e^{i\omega y}\,dy \\ \frac{\partial^2 u}{\partial y^2} = \frac{1}{2\pi}\int_{\infty}^{\infty} -\omega^2\hat{u}(x,\omega)e^{i\omega y}\,dy \end{gathered} \tag{2.59}$$

So that the Laplace's equation becomes:

$$\begin{aligned} \frac{\partial^2 \hat{u}}{\partial x^2} - \omega^2\hat{u} = 0 \qquad &\Rightarrow \hat{u} = A(\omega)e^{\omega x} + B(\omega)e^{-\omega x} \\ \Rightarrow u(x,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} &\left[A(\omega)e^{\omega x} + B(\omega)e^{-\omega x}\right] e^{i\omega y}\,dy \end{aligned} \tag{2.60}$$

where we can find the values for the coefficients A and B by applying the boundary conditions:

$$
\begin{aligned}
u(0,y) = g_1(y) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}[A(\omega)+B(\omega)]e^{i\omega y}\,d\omega \\
\frac{\partial u}{\partial x}(0,y) = g_2(y) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\omega[A(\omega)-B(\omega)]e^{i\omega y}\,dy \\
A(\omega)+B(\omega) = \hat{g}_1(y) &\qquad \omega[A(\omega)-B(\omega)] = \hat{g}_2(\omega) \\
A(\omega) = \frac{\hat{g}_1(\omega)+\hat{g}_2(\omega)/\omega}{2} &\qquad B(\omega) = \frac{\hat{g}_1(\omega)-\hat{g}_2(\omega)/\omega}{2}
\end{aligned}
\tag{2.61}
$$

And the derivative:

$$
\frac{\partial u}{\partial x} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\omega[A(\omega)e^{\omega x}-B(\omega)e^{\omega x}]e^{i\omega y}\,dy \tag{2.62}
$$

In general $A(\omega) \neq 0$, so there exists an exponentially amplified mode. This shows that the problem is ill-posed, to be well-posed, only one condition has to be specified as initial condition, to close the system a so called far field condition has to be specified. This implies that $A = 0$, but it still remains the $B \neq 0$ that shows that the solution is damped. If we introduce a perturbation into the domain it will be felt in the whole domain while being damped as one gets away from the perturbation source, the region of dependence and the region of influence are the entire domain.

We have some examples:

- if we inject a heating resistor in a pool, we will have a temperature increase in the whole pool, but it decreases exponentially.
- when a plane flies, we feel a small pressure difference at the ground. At the difference, for supersonic flights, we experience the sonic boom because there is no damping and we feel the region of influence of the perturbation.

2.4.2 Well posed problem for an elliptic system

We have already said that for the problem to be well-posed, one condition has to be specified at the initial boundary and the other at the other end of the domain. We can show that this is in fact like that for any arbitrary domain. Let's show that the boundary value problem $\Delta u = 0$ in the domain Ω with u specified on the whole domain boundary Γ is well posed.

Existence of the solution

This is beyond the scope of the course, just mention the Poisson formula:

$$
u = -\frac{1}{2\pi}\oint_{\Gamma}\left(\ln r\frac{\partial u}{\partial n} - u\frac{\partial \ln r}{\partial n}\right)d\Gamma \tag{2.63}
$$

where r is the distance between the solution evaluation point P and a running boundary point. If in addition we can define a harmonic function ν over Γ taking the same values as $\ln r$ on the boundary is known, then:

$$
u = -\frac{1}{2\pi}\oint_{\Gamma} u\left(\frac{\partial \nu}{\partial n} - \frac{\partial \ln r}{\partial n}\right)d\Gamma \tag{2.64}
$$

which proves the existence of the solution.

Solution uniqueness

Consider solutions u_1 and u_2, where $\nu = u_1 - u_2$ is the solution of the problem $\Delta\nu = 0$ with homogeneous boundary conditions. Let's show that $\nu = 0$:

$$\int_\Gamma \nu\Delta\nu \, d\Omega = \int_\Omega \nu\nabla\cdot(\nabla\nu) \, d\Omega = \int_\Omega \nabla\cdot(\nu\nabla\nu) \, d\Omega - \int_\Omega \nabla\nu\cdot\nabla\nu \, d\Omega = 0 \tag{2.65}$$

But Gauss tells that:

$$\int_\Omega \nabla\cdot(\nu\nabla\nu) \, d\Omega = \int_\Gamma \nu\frac{\partial\nu}{\partial n} d\Gamma = 0 \tag{2.66}$$

Because of the boundary conditions it is deduced that:

$$-\int_\Omega \nabla\nu\cdot\nabla\nu \, d\Omega = 0 \tag{2.67}$$

for this to be satisfied, $\nabla\nu = 0 \Rightarrow \nu = cst = 0$ to satisfy the homogeneous boundary conditions.

Continuity in the sense of Hadamard

This is verified directly by Poisson equation, if we perturb u infinitesimally the integral do so.

2.5 Parabolic equations

They are kind of degenerated hyperbolic equations since they have real characteristics but in a number lower than the number of unknowns. Let's consider the one unknown parabolic equation which admits two identical real characteristics:

$$\epsilon\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} + \frac{\partial u}{\partial t} = 0 \tag{2.68}$$

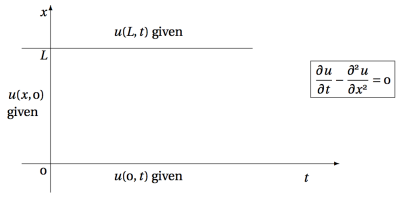

Figure 2.5

When $\epsilon > 0$, the equation is hyperbolic with characteristic speed $\pm 1/\sqrt{\epsilon}$, when $\epsilon \to 0$ the characteristic lines tend toward vertical lines in x, t plane. It can be deduced that the region of dependence and the region of influence are a half plane. A perturbation at point x_0, t_0 affects all the points $t > t_0$. Since the second derivative disappears, only one condition is to be imposed on the boundary and initial data as shown on the figure (evolution of the temperature in a metal bar).

2.6 Relation between differential problem type and discretized algebraic system structure

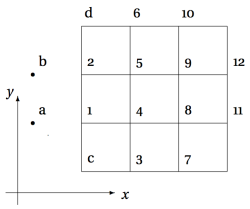

Figure 2.6

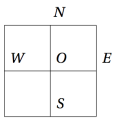

Figure 2.7

We will show that the type of differential equation to be solved (evolution or equilibrium) is related to the structure of the algebraic system resulting from the discretization. Consider the following equation, 2D small disturbance potential equation:

$$(1-M_\infty^2)\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=0 \qquad S^2-RT=M_\infty^2-1 \tag{2.69}$$

$1-M_\infty^2>0$ (hyperbolic) $M_\infty>1$, <0 (elliptic) $M_\infty<1$. We discretize by central differences on a very coarse 4x4 mesh with equal spacing in both directions $\Delta x=\Delta y=h$. If I consider a generic point O, We denote the neighbors by N W E S and the central finite difference expression of the general equation will be:

$$\begin{gathered}(1-M_\infty^2)\frac{\varphi_W+\varphi_E-2\varphi_O}{\Delta x^2}+\frac{\varphi_N+\varphi_S-2\varphi_O}{\Delta y^2}=0 \qquad \Delta x=\Delta y=h \\ \Rightarrow(1-M_\infty^2)[\varphi_W+\varphi_E-2\varphi_O]+[\varphi_N+\varphi_S-2\varphi_O]=0\end{gathered} \tag{2.70}$$

2.6.1 Subsonic flow $M_\infty=0$

If the equation is elliptic, we just have to give one boundary condition on the whole domain boundary. For simplicity let's prescribe φ (Dirichlet condition). We can build a matrix, characterizing a set of linear equations $A\varphi=b$:

$$A=\begin{pmatrix}
1 & & & & & & & & & & & & & & & \\
 & 1 & & & & & & & & & & & & & & \\
 & & 1 & & & & & & & & & & & & & \\
-1 & & -1 & 4 & -1 & & & -1 & & & & & & & & \\
 & -1 & & -1 & 4 & -1 & & & -1 & & & & & & & \\
 & & & & & 1 & & & & & & & & & & \\
 & & & & & & 1 & & & & & & & & & \\
 & & & -1 & & & -1 & 4 & -1 & & -1 & & & & & \\
 & & & & -1 & & & -1 & 4 & -1 & & -1 & & & & \\
 & & & & & & & & & 1 & & & & & & \\
 & & & & & & & & & & 1 & & & & & \\
 & & & & & & & & & & & 1 & & & & \\
 & & & & & & & & & & & & 1 & & & \\
 & & & & & & & & & & & & & 1 & & \\
 & & & & & & & & & & & & & & 1 & \\
 & & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Notice that that the structure is not an upper or lower triangular structure and requires efficient solving methods. This is an equilibrium problem, in structural analysis Gaussian elimination is used but in fluid dynamics it is too costly and we use iteration methods.

2.6.2 Supersonic flow $M_\infty=\sqrt{2}$

For supersonic case, we have an evolution problem and we have to specify $\phi, \frac{\partial\varphi}{\partial x}$ on the left boundary and one condition on the upper and lower boundaries, but no more at the outlet. Let's discretize the boundary points 1 and 2, we need external points a and b:

$$\begin{gathered}-\frac{\varphi_4-2\varphi_1+\varphi_a}{h^2}+\frac{\varphi_2-2\varphi_1+\varphi_c}{h^2}=0 \\ \Rightarrow \varphi_2+\varphi_c-\varphi_4-\varphi_a=0 \qquad \left.\frac{\partial\varphi}{\partial x}\right|_1=\frac{\varphi_4-\varphi_a}{2h}=u_1 \text{ (specified)}\end{gathered} \tag{2.71}$$

At this stage it is possible to eliminate the fictitious point a using the derivative, same is done for point 2 in the second equation:

$$-2\varphi_4+\varphi_2=-\varphi_c-2hu_1 \qquad -2\varphi_5+\varphi_1=-\varphi_d-2hu_2 \tag{2.72}$$

Grouping all together we again find a matrix:

$$A = \begin{pmatrix}
1 & & & & & & & & & & & \\
 & 1 & & & & & & & & & & \\
 & & 1 & & & & & & & & & \\
 & 1 & & -2 & & & & & & & & \\
1 & & & & -2 & & & & & & & \\
 & & & & & 1 & & & & & & \\
 & & & & & & 1 & & & & & \\
-1 & & 1 & & 1 & & -1 & & & & & \\
 & -1 & & 1 & & 1 & & -1 & & & & \\
 & & & & & & & & 1 & & & \\
 & & & -1 & & & 1 & & 1 & & -1 & \\
 & & & & -1 & & & 1 & & 1 & & -1
\end{pmatrix}$$

It is this time lower triangular and thus the solution is easily found by forward substitution. It is interesting to observe that the mathematical nature of the equation (marching in the zone of influence) is reflected in the method of solving (forward substitution). The problem of solving method does not arise here, the problem is the amplification of round off errors in the substitution process (stability of the method).

Chapter 3

Numerical methods for evolution problems - Stability

As seen previously, evolution problems concern is the stability of the forward substitution. We will introduce the subject on ordinary differential equations since after discretization a partial differential equation is transformed into a system of ordinary differential equations. We are going to consider model equations for hyperbolic and parabolic equations which will be respectively advection and diffusion equations:

$$\frac{\partial u}{\partial t} + \underbrace{a}_{\text{or } u}\frac{\partial u}{\partial x} = 0 \qquad \frac{\partial u}{\partial t} = \alpha\frac{\partial^2 u}{\partial x^2} \tag{3.1}$$

3.1 Consistency - stability - convergence

Convergence is a concept we have introduced before, the numerical method is convergent if $\lim_{h\to 0} ||u^h - u|| = 0$ when the number of parameters tends towards infinity, where u^h is the approximate numerical solution and u the exact solution and h is representative of the mesh size. For methods based on a functional representation (finite elements, spectral methods), convergence can be proven directly using functional analysis, see literature. When it is time dependent problem, we use finite differences for time integration (we don't need flexibility for time it is uniform). Finite differences convergence cannot be proven directly, but in contrasts is proven indirectly using the concept of consistency and stability.

3.1.1 Consistency

We say that a numerical method is consistent if $\lim_{h\to 0} ||D^h(u^h) - D(u)|| = 0$ when the mesh size tends towards 0, where $D(u)$ is the differential operator applied to exact solution and the other discrete approximation of the differential operator. That's nothing else but the truncation error. For example if the problem was:

$$\frac{\partial u}{\partial x} = f \qquad \frac{u_{i+1} - u_{i-1}}{2h} = f_i \qquad \underbrace{\Rightarrow}_{+} \frac{u_{i+1} - u_{i-1}}{2h} - \left(\frac{\partial u}{\partial x}\right)_i + f_i - f_i = TE = \mathcal{O}(\Delta x^2) \tag{3.2}$$

which shows that we have indeed the truncation error. For ODE the truncation error only depends on the mesh size $\mathcal{O}(h^p)$, so we need at least a first order accurate $p \leq 1$ method. For PDE, this depends on the mesh spacing in all variables under various combinations, and sometimes it is imposed to be consistent. Consider for example the advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{3.3}$$

The Lax-Friedrichs method gives:

$$\begin{aligned}
&\frac{u_i^{n+1} - \frac{1}{2}(u_{i-1}^n + u_{i+1}^n)}{\Delta t} + a\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0 \\
\text{but } &\begin{cases} u_{i+1}^n = u_i^n + \Delta x \left(\frac{\partial u}{\partial x}\right)_i^n + \frac{\Delta x^2}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_i^n + \frac{\Delta x^3}{6}\left.\frac{\partial^3 u}{\partial x^3}\right|_i^n \\ u_i^{n+1} = u_i^n + \Delta t \left(\frac{\partial u}{\partial t}\right)_i^n + \frac{\Delta t^2}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_i^n \end{cases} \\
\Rightarrow &\frac{\Delta t\left(\frac{\partial u}{\partial t}\right)_i^n + \frac{\Delta t^2}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_i^n - \frac{\Delta x^2}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_i^n + \dots}{\Delta t} + a\left(\left.\frac{\partial u}{\partial x}\right|_i^n + \frac{\Delta x^2}{3}\left.\frac{\partial^3 u}{\partial x^3}\right|_i^n + \dots\right) \\
&= \left(\frac{\partial u}{\partial t}\right)_i^n + a\left(\frac{\partial u}{\partial x}\right)_i^n + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} - \frac{\Delta x^2}{2\Delta t}\frac{\partial^2 u}{\partial x^2} + \dots
\end{aligned} \tag{3.4}$$

We see that when $\Delta t \to 0$, the first term is 0 but the second one is 0 only if $\lim_{\Delta t, \Delta x \to 0}(\Delta x^2/\Delta t) = 0$.

3.1.2 Stability

Following Lax's definition, a numerical method is said to be stable if the numerical solution in a value T of the evolution variable t remains bounded as the mesh size $\Delta t \to 0$. For ODE:

$$\lim_{\Delta t \to 0, n\to\infty, n\Delta t = T} u^n \text{ exists} \tag{3.5}$$

3.1.3 Convergence

A numerical method is said to be convergent if:

$$\lim_{h\to 0}(u - u^h) = 0 \tag{3.6}$$

We prove convergence from consistency and stability using Lax's equivalence theorem:

Lax's equivalence theorem

For a well posed initial value problem and a consistent discretization method, stability is the necessary and sufficient condition for convergence:

$$\text{Consistency} + \text{Stability} \Leftrightarrow \text{Convergence}$$

3.2 Spectrum of the space discretization - Fourrier analysis

If we start from a PDE and we discretize all the variables except the evolution variables, we get a system of ODE's:

$$\frac{dU}{dt} = F(U) \tag{3.7}$$

If the PDE is linear, and if the discretization formula is linear, then this system is linear as well:

$$\frac{dU}{dt} = SU + Q \tag{3.8}$$

where U is an n vector and S is an $n \times n$ matrix (n is the number of degrees of freedom). So far all the discretization methods we discussed, have been linear, for example:

$$\frac{\partial u}{\partial x} \approx \frac{u_{i+1} + u_{i-1}}{2\Delta x} \tag{3.9}$$

because if we replace u by a combination of u and v we will have a linear combination. It turns out that for hyperbolic problems and in particular for non-linear ones, linear discretization will produce purely oscillatory solutions, except in first order. It has been discovered that it was possible to go beyond first order discretization without oscillatory solution but this is not viewed in the frame of this course.

If we assume that the matrix S has a complete set of eigenvalues and eigenvectors (real or complex) then we know that the matrix can be diagonalized such that $LS = \Lambda L$:

$$L\frac{dU}{dt} - LSU = LQ \Leftrightarrow L\frac{dU}{dt} - \Lambda LU = LQ \qquad \Rightarrow \frac{dw_i}{dt} - \lambda^{(i)} w_i = (LQ)_i \tag{3.10}$$

which gives a set of uncoupled equations, model equations. This allows to separately study the stability of time integration for each mode, that depends on the eigenvalue spectrum of the matrix S resulting from the space discretization of the differential operator. Finding the space discretization is complex because depends on the PDE and the discretization. However, some analytic results have been obtained for some model problems.

3.2.1 Spectrum of the central space discretization of the diffusion equation

The equation we want to look at is the model of parabolic equations (diffusion equation):

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \tag{3.11}$$

where α is the diffusivity coefficient $[L^2T^{-1}]$. The evolution parameter is t, we want to discretize the equation by central finite differences on a mesh of $N+1$ points in the interval $0 \leq x \leq L$ (so $\Delta x = L/N$). The discretization of an internal point ($1 \leq i \leq N-1$) is:

$$\frac{du_i}{dt} = \alpha \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \tag{3.12}$$

We will consider 3 problems varying from boundary conditions. We have to impose an initial condition on one boundary at left and right boundary.

Dirichlet conditions

Here $u(0,t) = u_0(t) = a, u(N,t) = u_N(t) = b$ are specified. These can easely be eliminated by having the following system for the remaining $N-1$ points:

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \frac{\alpha}{\Delta x^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} + \begin{pmatrix} \frac{\alpha u_0}{\Delta x^2} \\ 0 \\ \vdots \\ 0 \\ \frac{\alpha u_N}{\Delta x^2} \end{pmatrix} \tag{3.13}$$

The eigenvalues are (not detailed in the course):

$$\lambda^{(j)} = \frac{-4\alpha}{\Delta x^2}\sin^2\left(\frac{\pi j}{2N}\right) \qquad j = 1, \ldots, N-1 \tag{3.14}$$

Notice that they are all negative reals and they span a very large range, the first Fourrier mode corresponds to a wave of wavelength $2L$:

$$\begin{aligned}\lambda_1 &= -\frac{4\alpha}{\Delta x^2}\sin^2\frac{\pi}{2N} \approx -\frac{4\alpha}{\Delta x^2}\left(\frac{\pi}{2N}\right)^2 = -\frac{\alpha}{L^2}\pi^2 \\ \lambda_{N-1} &= -\frac{4\alpha}{\Delta x^2}\sin^2\frac{\pi(N-1)}{2N} \approx \frac{-4\alpha}{\Delta x^2}\end{aligned} \tag{3.15}$$

Neumann conditions at the left end and Dirichlet at the right end

$\left.\frac{\partial u}{\partial x}\right|_0 = a, u_N$ specified, we have to express the discretization of the left boundary point (i = 0). This can be done like:

$$\begin{aligned}\frac{du_0}{dt} &= \frac{\alpha}{\Delta x^2}(u_1 - 2u_0 + u_{-1}) = \frac{\alpha}{\Delta x^2}(2u_1 - 2u_0) - \frac{2\alpha a}{\Delta x} \\ \left.\frac{\partial u}{\partial x}\right|_0 &= \frac{u_1 - u_{-1}}{2\Delta x} = a \Rightarrow u_{-1} = u_1 - 2a\Delta x\end{aligned} \tag{3.16}$$

So that here we have N unknowns, and only the first line changes:

$$\frac{d}{dt}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1}\end{pmatrix} = \frac{\alpha}{\Delta x^2}\begin{pmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2\end{pmatrix}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1}\end{pmatrix} + \begin{pmatrix} -\frac{2\alpha a}{\Delta x} \\ 0 \\ \vdots \\ 0 \\ \frac{\alpha u_N}{\Delta x^2}\end{pmatrix} \tag{3.17}$$

The eigenvalues are:

$$\lambda^{(j)} = \frac{-4\alpha}{\Delta x^2}\sin^2\left(\frac{(2j-1)\pi}{4N}\right) \qquad j = 1, \ldots, N \tag{3.18}$$

The first Fourrier mode is now a wave with wavelength $= 4L$:

$$\begin{aligned}\lambda_1 &= -\frac{4\alpha}{\Delta x^2}\sin^2\frac{\pi}{4N} \approx -\frac{4\alpha}{\Delta x^2}\left(\frac{\pi}{4N}\right)^2 = -\frac{\alpha}{4L^2}\pi^2 \\ \lambda_N &= -\frac{4\alpha}{\Delta x^2}\sin^2\frac{\pi(2N-1)}{4N} \approx \frac{-4\alpha}{\Delta x^2}\end{aligned} \tag{3.19}$$

Periodic boundary conditions

Now $u_0 = u_N$ and the system is

$$\frac{d}{dt}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1}\end{pmatrix} = \frac{\alpha}{\Delta x^2}\begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2\end{pmatrix}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1}\end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0\end{pmatrix} \tag{3.20}$$

The eigenvalues are:

$$\lambda^{(j)} = \frac{-4\alpha}{\Delta x^2}\sin^2\left(\frac{\pi j}{N}\right) \qquad j = 0, \ldots, N-1 \tag{3.21}$$

Since the system has N unknowns, it has N eigenvalues, the first one is zero since the system matrix is singular. The first to be non zero is the first Fourrier mode of wavelength L:

$$\begin{aligned} \lambda_0 &= 0 \\ \lambda_1 &= -\frac{4\alpha}{\Delta x^2}\sin^2\frac{\pi}{N} \approx \frac{-4\alpha}{L^2}\pi^2 \\ \lambda_{\frac{N-1}{2}} &\approx -\frac{4\alpha}{\Delta x^2} \end{aligned} \tag{3.22}$$

The largest modulus eigenvalue is the one for $j = (N-1)/2$. Now we can decompose the u variable in its time and space functions, the diffusion equation becomes:

$$\frac{\partial u}{\partial t} = \alpha\frac{\partial^2 u}{\partial x^2} \qquad u = \hat{u}(t)g(x) \qquad g(x)\frac{d\hat{u}}{dt} = \alpha\hat{u}(t)g^{"}(x) \qquad \Rightarrow \frac{d\hat{u}}{\hat{u}dt} = \alpha\frac{g^{"}}{g} \tag{3.23}$$

and if we define g such that:

$$g^{"} + k^2 g = 0 \qquad \Rightarrow \frac{d\hat{u}}{\hat{u}dt} = \alpha\frac{g^{"}}{g} = cst - \alpha k^2 \qquad \Leftrightarrow \frac{d\hat{u}}{dt} + \alpha k^2\hat{u} = 0 \tag{3.24}$$

where

$$\begin{gathered} g = A\cos kx + B\sin kx \qquad g(0) = g(L) = 0 \Rightarrow A = 0 \\ B\sin kL = 0 \Rightarrow kL = m\pi \Rightarrow k = \frac{m\pi}{L} \end{gathered} \tag{3.25}$$

Fourier analysis

As the eigenvalue computations are difficult, a method consist in approximating the spectrum based on Fourier analysis. Remember that a running discrete equation at inner point was given by:

$$\frac{du_i}{dt} = \alpha\frac{u_{i+1} - 2u_i + u_{i+1}}{\Delta x^2} \tag{3.26}$$

The Fourier analysis is to assume a periodic solution in space:

$$u(x,t) = \hat{u}(t)e^{ikx} \qquad u_i(t) = u(x_i,t) = \hat{u}(t)e^{ikx_i} \qquad \begin{aligned} u_{i\pm1}(t) &= u(x_{i\pm1},t) = \hat{u}(t)e^{ikx_{i\pm1}} \\ &= \hat{u}(t)e^{ik(x_i\pm\Delta x)} \\ &= u_i e^{\pm ik\Delta x \equiv \pm i\eta} \end{aligned} \tag{3.27}$$

where η is non-dimensional and is the **reduced wave number**. Inserting these results in the discretized equation we get:

$$\frac{du_i}{dt} = \alpha\frac{e^{i\eta} - 2 + e^{-i\eta}}{\Delta x^2}u_i = \frac{2\alpha}{\Delta x^2}\underbrace{(\cos\eta - 1)}_{1-2\sin\frac{\eta}{2}}u_i = \underbrace{-\frac{4\alpha}{\Delta x^2}\sin^2\frac{\eta}{2}}_{\text{Fourier footprint } \lambda} u_i \tag{3.28}$$

The locus of all possible values of λ as a function of the reduced wavenumber is called the **Fourrier footprint** of the discretized equation. Since $-1 \leq \sin\frac{\eta}{2} \leq 1$, in this case it is located on the negative real axis between 0 and $-4\frac{\alpha}{\Delta x^2}$ and is indeed an approximation of what we obtained with the previous boundary condition discussion.

3.3 Spectra of various discretization of the advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{3.29}$$

which will be solved in the same interval 0 to L with $N+1$ points with $\Delta x = L/N$.

Backward upwind space-discretization

It means that when we make the discretization in space in an interior point we have:

$$\frac{du_i}{dt} + a\frac{u_i - u_{i-1}}{\Delta x} = 0 \tag{3.30}$$

Accordingly to Section 2.3.2, with $a > 0$ only a boundary condition in $x = 0$ has to be imposed $u(0,t) = u_0(t) = g(t)$, such that the matrix form as previously gives:

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = -\frac{a}{\Delta x}\begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} + \begin{pmatrix} \frac{a}{\Delta x}g(t) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{3.31}$$

Since the matrix is lower triangular, the eigenvalues are on the the diagonal elements:

$$\lambda^{(j)} = -\frac{a}{\Delta x} \qquad j = 1, \ldots, N \tag{3.32}$$

They are all negative and real → well posed problem and discretization. If we express in terms of modal expression using the λ, we have: $\frac{dw_j}{dt} = \frac{-a}{\Delta x}w_j \rightarrow w_j = w_{j0}e^{-\frac{a}{\Delta x}t}$. Looking to the exponential term, we observe that the space discretization in this case has provided damping to the solution. This is an inaccuracy in the method because we have seen in chapter 2 that the solution is a wave propagation without damping. In the case of $a > 0$, the damping transforms into amplification and this is even worse (positive eigenvalue so ill-defined problem).

Fourier analysis

Let's estimate the spectrum. When we do a Fourier analysis we assume the solution to be periodic:

$$u_{i\pm1} = u_i e^{\pm i\eta} \qquad \eta = k\Delta x \qquad u = \hat{u}(t)e^{ikx} \tag{3.33}$$

So that the discretization becomes:

$$\frac{du_i}{dt} + \underbrace{a\frac{1 - e^{-I\eta}}{\Delta x}}_{\lambda} u_i = 0 \tag{3.34}$$

If we try to draw this λ, we begin with a circle $e^{-I\eta}$ then transpose to the right $1 - e^{-I\eta}$ and then we change the direction, the radius is $a/\Delta x$ centered in $-a/\Delta x$. The form of the footprint is completely different and this is due to the boundary condition. If we have periodic boundary conditions:

$$\frac{d}{dt}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = -\frac{a}{\Delta x}\begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix}\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{3.35}$$

with the eigenvalues $\lambda_j = \frac{-a}{\Delta x}(1 - e^{-I2\pi j/N}) \quad j = 0, \ldots, N-1$. This method allows to prove that the forward discretization is inappropriate for this method since the Fourier footprint is a circle entirely situated on the right half plane.

Central discretization

For an interior point we have now:

$$\frac{du_i}{dt} + a\frac{u_{i+1} - u_{i-1}}{2\Delta x} = 0 \tag{3.36}$$

The problem is that for the last point N, we need a point on the right which does not exist, so we replace by the backward discretization at point N:

$$\frac{du_N}{dt} = -\frac{a}{\Delta x}(u_N - u_{N-1}) \tag{3.37}$$

The matrix system we get is:

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = -\frac{a}{\Delta x}\begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} + \begin{pmatrix} \frac{a}{2\Delta x}g(t) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{3.38}$$

If we make a Fourier analysis of that we can find that:

$$\frac{du_i}{dt} + a\frac{u_{i+1} - u_{i-1}}{2\Delta x} = 0 \qquad u_{i\pm1} = u_i e^{\pm I\eta} \tag{3.39}$$

$$\frac{du_i}{dt} + a\frac{e^{I\eta} - e^{-I\eta}}{2\Delta x}u_i = 0 \qquad \frac{du_i}{dt} = -\frac{2Ia\sin\eta}{2\Delta x}u_i = -I\frac{a}{\Delta x}u_i\sin\eta \tag{3.40}$$

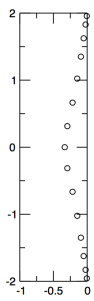

Figure 3.1

The eigenvalues of the matrix cannot be found analyticlly but havec to be computed numerically, the result is shown on the figure. We can observe that all the eigenvalues are in the negative real half plane but much closer to the imaginary axis than the backward method. This implies that the numerical damping is much smaller. When looking to the Fourier analysis, we remark that the footprint is purely imaginary. The important conclusion is that we see a fundamental different behavior between the diffusion and advective, advection is essentially an imaginary footprint with a possible negative real part while the other has real negative footprint.

3.4 Stability of time-integration schemes for ODE

3.4.1 Definition — examples

We will study the stability of time integration with the following linear homogeneous test problem:

$$\frac{du}{dt} = f(u,t) \qquad \Rightarrow \frac{du}{dt} = \underbrace{q}_{\lambda} u \qquad u(0) = 1 \tag{3.41}$$

where q is complex $q = \sigma + iw$ and the solution $u = e^{qt}$ corresponds to a stable behavior if $\mathcal{R}(q) \leq 0$. When discretization by finite difference is applied on it, the equation can be cast in the form:

$$u^{n+1} = g(q, \Delta t)u^n \qquad \Rightarrow u^{n+1} = [g(q, \Delta t)]^n u^1 \tag{3.42}$$

where g is generally complex and is called the **amplification factor**. The stability condition then requires that $[g(q, \Delta t)]$ must be uniformly bounded for $0 < \Delta t < t, 0 \leq n\Delta t \leq T$. A necessary condition is that $|g(q, \Delta t)| \leq 1 + \mathcal{O}(\Delta t)$ and in particular that $|g(q, 0)| \leq 1|$. Stable if the numerical solution remains bounded when the number of steps n goes to infinity and the time step size goes to 0. There are two examples on p.81, here is the second one more interesting:

EXAMPLE: Let us show that the forward finite difference discretization is stable:

$$\frac{dun}{dt} \approx \frac{u^{n+1} - u^n}{\Delta t} = qu^n \qquad u^{n+1} = (1 + q\Delta t)u^n \qquad g = 1 + q\Delta t = 1 + \mathcal{O}(\Delta t) \tag{3.43}$$

Let us use a centered finite difference discretization of the time derivative du/dt (two step explicit mid-point method)

$$\frac{du}{dt} = qu \Rightarrow \frac{u^{n+1} - u^{n-1}}{2\Delta t} = qu^n \qquad \Rightarrow u^{n+1} = u^{n-1} + 2\Delta t q u^n \tag{3.44}$$

Let's look for g such that $u^{n+1} = gu^n = g^2u^{n-1}$, replacing everything we get:

$$(g^2 - 2q\Delta t g - 1)u^{n-1} = 0 \qquad \Rightarrow g = q\Delta \pm \sqrt{1 + (q\Delta t)^2} \tag{3.45}$$

This method is thus stable since $|g| \leq 1 + \mathcal{O}(\Delta t)$ and in particular $g(q, 0) = \pm 1$.

3.4.2 Weak (in)stability

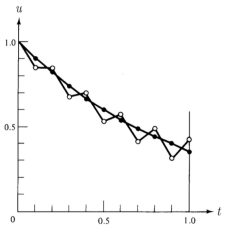

Figure 3.2

Consider the centered discretization for the test problem with $q = -1, \Delta t = 0.1$. Since it is a two step method, one should provide two initial values. One is provided by the initial condition $u^0 = 1$. To compute the first point, we will apply the forward discretization (cause centered need previous point):

$$\frac{u_1 - u_0}{\Delta t} = qu_0 \qquad \rightarrow u_1 = (1 + q\Delta t)u_0 = 0.9 \tag{3.46}$$

which we have shown it is stable as well. This is a good approx since $u(-1, 0.1) = e^{-0.1} = 0.905$. The result of the numerical computation is shown on the figure for $u^1 = 0.9$ and for $u^1 = 0.85$. One can observe that the small perturbation on u^1 gives rise to amplifying oscillations. As small as the initial perturbation can be, there will always be amplification. This is called **weak instability** and is unacceptable.

Let's look for the reason, we have seen that we can have two values for g (solution of quadratic equation) with centered discretization. An expression of the form the following form is solution of the difference equation:

$$u^n = c_1 g_1 + c \tag{3.47}$$

where c_1 and c_2 are found by the boundary conditions:

$$u^0 = c_1 + c_2 \qquad u^1 = c_1 g_1 + c_2 g_2 \tag{3.48}$$

Let's now look at what happens when we rise the coefficient by the exponent n, from previous analysis we know that for $q = -1$:

$$g_{1,2} = -\Delta t \pm \sqrt{1 + \Delta t^2} \tag{3.49}$$

where we remark that $|g_1| < 1$ and $|g_2| > 1, g_2 < 0$. When we make the Taylor expansion of these, we refind the Taylor expansion of the exponentials:

$$\begin{aligned}
g_1 &= 1 - \Delta t + \frac{\Delta^2}{2} + \mathcal{O}(\Delta t^3) = e^{-\Delta t} + \mathcal{O}(\Delta t^3) = e^{-\Delta t + \mathcal{O}(\Delta t^3)} \\
g_2 &= -1 - \Delta t - \frac{\Delta^2}{2} - \mathcal{O}(\Delta t^3) = (-1)(e^{\Delta t} + \mathcal{O}(\Delta t^3)) = (-1)e^{\Delta t + \mathcal{O}(\Delta t^3)} \\
g_1^n &= e^{-n\Delta t} e^{\mathcal{O}(n\Delta t^3)} = e^{-t} e^{\mathcal{O}(n\Delta t^2)} \\
g_2^n &= (-1)^n e^{n\Delta t} e^{\mathcal{O}(n\Delta t^3)} = (-1)^n e^{t} e^{\mathcal{O}(n\Delta t^2)}
\end{aligned} \tag{3.50}$$

where we see that the problem comes from g_2 which has an increasing exponential and changing sine every time step n. The term $c_2 g_2$ has no relation with the exact solution and is a numerical artefact. It is impossible to have $c_2 = 0$ because there will always be round-off errors as small they can be. The problem with the stability definition is that it limits itself to $\Delta t \to 0$ and not finite one.

3.4.3 Region of (absolute stability)

In that definition, the region of stability of numerical algorithm for integrating an ODE is defined as the set of values of the complex variable $z = q\Delta t$ such that the sequence u^n of numerical values remains bounded as $n \to \infty$ (no more $\Delta \to 0$). This is equivalent to stating that the origin $z = q\Delta t = 0$ lies in the region of absolute stability.

Forward Euler method

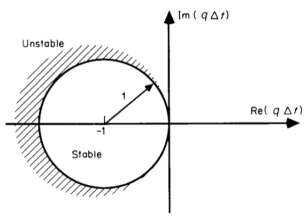

Figure 3.3

Let's begin with this first example:

$$\frac{u^{n+1} - u^n}{\Delta t} = q u^n \qquad \Rightarrow g = 1 + q\Delta t = 1 + z \tag{3.51}$$

The region of stability is $|1 + z| \leq 1 \to$ we have a circle centered at $z = -1$ as shown on the figure. With $q = -1$ we find out that the condition is $\Delta \leq 2$. This is not a severe restriction, Δt is thus more limited for accuracy condition than for stability.

Central finite difference method

The amplification factor was:

$$g^2 - 2q\Delta t g \underbrace{-1}_{g_1 g_2} = 0 \qquad \Rightarrow g = q\Delta t \pm \sqrt{1 + (q\Delta t)^2} \tag{3.52}$$

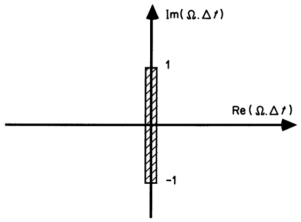

Figure 3.4

Since we have a second order equation we see that $|g_1 g_2| = 1$ and the only possibility to have $|g_{1,2} \leq 1|$ is to have $|g_1| = |g_2| = 1 \rightarrow g = e^{i\alpha}$. If we isolate z in the second order equation:

$$z = q\Delta t = \frac{g - 1/g}{2} = \frac{e^{i\alpha} - e^{-i\alpha}}{2} = 2i \sin\alpha \tag{3.53}$$

From which we deduce that the region of stability is the imaginary segment $[-i, i]$. It is thus not surprising that our previous computation with $q = -1, \Delta t = 0.1 \rightarrow z = -0.1$ is unstable since it is outside of the region of stability.

3.4.4 Stiff problems

The previous case was already a bit restrictive, let's see another example extremely restrictive:

$$\frac{du}{dt} = 100(\sin t - u) \qquad u(0) = 0 \tag{3.54}$$

which is a problem composed by a forcing term $\sin t$ and a homogeneous term u. The solution is a linear combination of the forced periodic response and a transitory term:

$$u(t) = \frac{\sin t - 0.01 \cos t + 0.01 e^{-100t}}{1.0001} \tag{3.55}$$

Due to the relatively small coefficient, the transient dies very quickly and we can keep interest on the forced response of period 2π only. One can thus choose a time step of let's say $\Delta t = 2\pi/20 \approx 0.3$. Using the Runge-Kutta method on matlab one gets the following results:

Δt	0.015	0.020	0.025	0.030
Number of steps	200	150	120	100
$u(3)$	0.151004	0.150996	0.150943	$6.7\ 10^{11}$

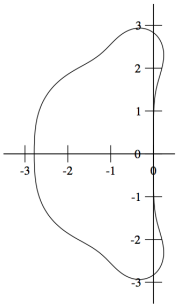

Figure 3.5

We see that for a very small time step as 0.03 the result blows up. This is due to the homogeneous term. Indeed, if we neglect the periodic term, we have $\frac{du}{dt} = -100u$ so that $q = -100$ and the region of stability on the figure shows that we must have $q\Delta t \leq 2.8 \rightarrow \Delta t \leq 0.028$. The limiting term is thus the homogeneous one. The difficulty comes from the coexistence of two phenomena with very different time scales and it is the shortest time scale that determines the maximum allowable time step. Problems with very different time scales phenomena are called **stiff problems** and are common in fluid mechanics. For example the discretization of the one-dimensional heat equation gave $\lambda_1 = \mathcal{O}(\alpha/L^2)$ and $\lambda_N = \mathcal{O}(\alpha/\Delta x^2)$, the ratio of the two time scale is thus $\mathcal{O}(L/\Delta x^2)$ which becomes very large for fine meshes. We want to find something stable independently of the time step, this is called **absolute stability** or **A-stability**.

3.4.5 Absolute stability

If we translate what we have introduce at the end of previous section into maths: a homogeneous modal problem such as the test problem $du/dt = qu$ is stable if $\mathcal{R}(q) \leq 0$. Therefore the set of value of $q\Delta t$ corresponding to stable is the left half plane. We remark that the Forward Euler and the Runge-Kutta methods are not A-sable. What about the backward Euler?

$$\frac{u^n - u_{n-1}}{\Delta t} = qu^n \qquad \Rightarrow u^n = \frac{1}{1 - q\Delta t} u^{n-1} \tag{3.56}$$

the region of stability requires thus that $|1 - q\Delta t| < 1$ which corresponds to the whole space without the region in a circle in the right half plane. The method is A-stable. However, this method requires to solve an equation in order to have the point u^n while the others not. In case of non-linear equation or system of equations the solving becomes more difficult. Methods like that are called **implicit** while methods like forward where we can compute the point directly using the previous one are called **explicit**. With implicit methods the set-up will be more complex and costly in computing time. A theorem by Dahlquist states that:

- An A-stable method must be implicit;
- An A-stable method has an accuracy of order $p \leq 2$.

It is thus not possible to have explicit A-stable methods. It is however possible to have methods with higher accuracy order by extending the class of methods under consideration (implicit Runge-Kutta method).

For stiff problems one will frequently choose among the simplest second order A-stable methods like the trapezoidal one:

$$\frac{u^{n+1} - u^n}{\Delta t} = \frac{q}{2}(u^n + u^{n+1}) \tag{3.57}$$

It can be shown that it is the method with the lowest truncation error. In summary, the gain in time step is not compensated by the cost in memory. On the other hand, extremely stiff problems like viscous and reactive flows requires the use of implicit methods.