Mémoire de fin d'études

Correction Finite-Volume Subcells pour schémas Galerkin Discontinu. Application aux équations Shallow-Water.

Sacha Cardonna

Juillet 2023

Encadré par Fabien Marche & François Vilar

Institut Montpelliérain Alexander Grothendieck
Département de Mathématiques – Université de Montpellier

Introduction

« Si vous touchez aux mathématiques, vous ne devez être ni pressés, ni cupides, fussiez-vous roi ou reine. »
Euclide d'Alexandrie (330 av. J.-C. – 270 av. J.-C.)

Ce mémoire est un résumé du travail que j'ai effectué à l'Institut Montpelliérain Alexander Grothendieck dans le cadre de mon stage de fin d'études. Notre document porte ici sur l'étude des équations aux dérivées partielles, en plus particulier l'étude de méthodes numériques d'ordres élevées, et leur application à des équations issues de la mécanique des fluides en eaux peu profondes.

Le premier chapitre introduit aux lecteurs les différentes méthodes sur lesquelles nous avons travaillé : y sont présentés leurs intérêts, leurs qualités mais aussi leurs limites. La seconde partie met en place les éléments mathématiques nécessaires à la compréhension correcte des chapitres suivants, qui concernent l'application des méthodes de stabilisation de François Vilar, introduites dans [4], sur une loi de conservation scalaire puis sur les équations de Saint-Venant. Sont également présent deux annexes : la première a été rédigée au début du stage à l'époque où je m'initiais aux méthodes Galerkin Discontinu, et la seconde est un travail théorique portant sur les équations de Saint-Venant et leur obtention à partir des équations d'Euler.

J'ai la chance de pouvoir considérer cet humble travail comme le début de ma vie de chercheur, car j'ai l'opportunité de continuer sur ce sujet-là en thèse. Je m'excuse donc d'avance si ce manuscrit est incomplet par moment et/ou développe des pistes non exploités par la suite ; elles le seront je l'espère pendant les trois prochaines années.

Même si je n'ai passé que quatre mois à leur côté, j'ai déjà une immense reconnaissance pour mes trois encadrants. En effet, j'ai rajouté au duo initial Ali Haidar, ancien thésard à l'IMAG, actuellement en post-doc à Nice, qui a pris de son temps pour m'aider alors qu'il n'en avait nullement l'obligation. Merci Ali, et j'espère qu'on va pouvoir continuer à collaborer ensemble.
Puis évidemment je souhaite terminer en remercier Fabien et François ; merci pour votre gentillesse, votre temps, et votre confiance.
Mes premiers pas dans le monde de la recherche s'inscrivent ici, aux côtés de trois scientifiques brillants et bienveillants, et j'ai hâte d'apprendre à marcher à leurs côtés *avant d'apprendre à courir*.

Contents

Section 1

Notions & motivations.

In the following section, we provide a brief description of the various components and techniques used to describe our numerical approach in this work.

1.1 Shallow-Water equations.

The *Shallow-Water* equations are a collection of partial differential equations that describe the behavior of fluids in shallow areas such as rivers, lakes, and coastal areas. Mathematicians, engineers, and scientists are all interested to them because they provide a fundamental framework for understanding fluid dynamics in a wide range of practical applications. SW equations were developed in the mid-nineteenth century by mathematicians and physicists who wanted to understand the behavior of water waves, obtained by deriving the full *Navier-Stokes* equations (see Appendice B), which describe fluid motion in general. The fluid was simplified by assuming that it is incompressible and inviscid, and that its depth is much smaller than its horizontal extent.

Given a smooth parametrization of the topography $b : \mathbb{R} \to \mathbb{R}$, denoting by H the water height, u the horizontal velocity and $q = Hu$ the horizontal discharge, the *pre-balanced* Nonlinear Shallow-Water equations may be written as follows :

$$\partial_t \mathbf{v} + \partial_x \mathbf{F}(\mathbf{v}, b) = \mathbf{B}(\mathbf{v}, \partial_x b),$$

where $\mathbf{v} : \mathbb{R} \times \mathbb{R}_+ \to \Theta$ gathers the flow's conservative variables and is assumed to take values in the convex and open set

$$\Theta = \{(\eta, q) \in \mathbb{R}^2 \mid H := \eta - b \geq 0\},$$

where $\mathbf{F} : \Theta \times \mathbb{R} \to \mathbb{R}^2$ is the nonlinear flux function and $\mathbf{B} : \Theta \times \mathbb{R} \to \mathbb{R}^2$ is the topography source term, defined as follows

$$\mathbf{v} = (\eta, q)^T, \qquad \mathbf{F}(\mathbf{v}, b) = \left(q, uq + \frac{1}{2} g\eta(\eta - 2b)\right)^T, \qquad \mathbf{B}(\mathbf{v}, \partial_x b) = (0, -g\eta\partial_x b)^T.$$

Those equations are hyperbolic, which means that information travels at a finite speed through the fluid, and nonlinear, which means that the fluid's behavior can be highly complex and difficult to predict. In addition, the equations are also conservative, which means that the total amount of fluid in a given region is constant over time.

SW equations have a wide range of practical applications in various fields, for example :

- In civil engineering, the equations are used to design and optimize hydraulic structures such as dams, canals, and water treatment plants. They are also used to study the effects of floods, tsunamis, and other extreme weather events on coastal infrastructure ;
- In environmental science, those equations are used to model the behavior of water and sediment in rivers, lakes, and estuaries. This allows scientists to study the impact of pollution, erosion, and other environmental factors on aquatic ecosystems. The equations are also used to study the behavior of ocean currents and their impact on global climate patterns ;
- In the energy sector, they can be used to model the behavior of fluid flows in oil and gas pipelines. This allows engineers to design more efficient and reliable pipelines that can transport fluids over long distances. The equations are also used to model the behavior of wind and tidal energy systems, which are becoming increasingly important sources of renewable energy.

Figure 1.1: *The Great Wave of Kanagawa*, Hokusai, 1830.

To summarize, the Shallow-Water equations represent a fundamental framework for understanding fluid dynamics in shallow regions, and they have practical applications in a wide range of fields. They are a testament to the power of mathematical modeling and its ability to provide insights into complex physical systems. As such, they represent an important area of research and development that will continue to be of great significance in the years to come. The curious reader will find information about how we can derive *Euler*'s equations to obtain SW in the appendice (B).

1.2 Discontinuous Galerkin methods.

The *Discontinuous Galerkin* (dG) method is a numerical scheme for solving partial differential equations. It was first introduced by Reed and Hill in 1973, and has since become a popular method for solving a wide range of problems, from fluid dynamics to electromagnetics. The dG method is based on the *Galerkin* method, which involves approximating a solution to a PDE as a linear combination of basis functions. However, unlike the continuous Galerkin method, which uses continuous basis functions, the dG method uses discontinuous basis functions. This allows for a more flexible and accurate approximation of the solution, particularly in areas with high gradients or shocks.

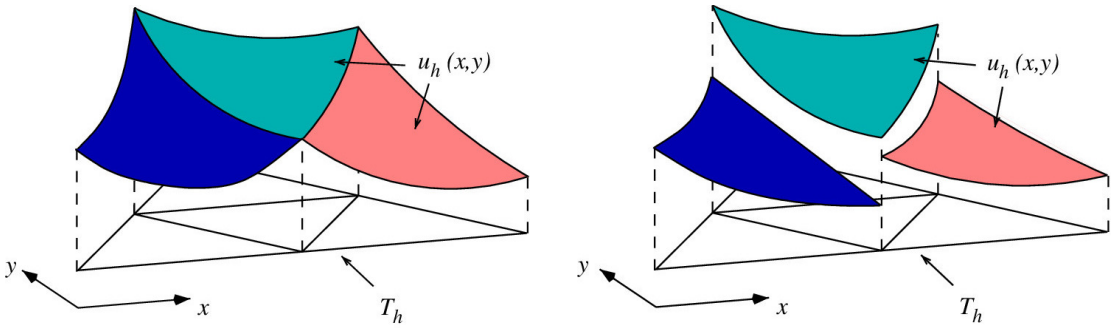


Figure 1.2: Differences between continuous and discontinuous Galerkin methods.

One of the main interests of the dG method is its ability to handle complex geometries and domains with irregular boundaries. This is because the method is naturally suited to handling non-uniform meshes and allows for the use of unstructured grids. The dG method is also well-suited to handle problems with multiple scales, such as those found in fluid dynamics or electromagnetism. Compared to other numerical methods, such as finite difference and finite element methods, the dG method has several advantages. For one, the dG method is more accurate and robust than other methods in areas with strong discontinuities or singularities.
This is because it can accurately capture the solution in these areas, whereas other methods may require finer mesh resolutions or more complex formulations. Another advantage of this method is its ability to handle conservation laws. The dG method naturally conserves mass, momentum, and energy, which is important for many applications, such as fluid dynamics and electromagnetism. In contrast, other methods may require additional stabilization techniques to enforce conservation.

1.3 Finite-Volume local subcell correction.

While the dG method has several advantages over other numerical methods, such as the *Finite-Volume* (FV) method, it also has some drawbacks that make it less robust in certain scenarios.

One of the main disadvantages of the dG method is its difficulty in handling strong shocks and discontinuities. This is because the method relies on discontinuous basis functions, which can lead to numerical oscillations and instability in the presence of too strong gradients. In contrast, the FV method uses piecewise constant reconstructions, which are better suited to capturing shocks and discontinuities.

Another drawback of the dG method is its computational expense. The dG method can be computationally expensive, particularly for high-order methods or complex geometries, due to the need for a large number of degrees of freedom and the cost of computing numerical fluxes at the element interfaces. In contrast, the FV method is generally more computationally efficient, particularly for lower-order methods and simpler geometries. Additionally, the dG method requires careful treatment of numerical fluxes at the interfaces between elements to ensure accuracy and stability. This can be particularly challenging in complex geometries or in the presence of strong shocks or discontinuities. In contrast, the FV method typically relies on simple numerical fluxes that are easy to implement and more robust in these scenarios.

Another challenging problem, focusing on the NSW equations, is the preservation of the set of admissible states (2.2) at the discrete level, which is closely related to the issue of the occurrence and propagation of wet/dry fronts that may occur in dam-breaks, flood-waves, or run-up over coastal shores. As a result, while maintaining the water-height positivity at the discrete level is a minimal nonlinear stability requirement, this is clearly a difficult task when high-order polynomials are used within mesh elements and standard (non-stabilized) dG methods may produce negative values for the water-height H in the vicinity of dry areas. In general, robustness issues may be among the most significant remaining challenges for the use of high-order methods in realistic problems in many domains of application, and in recent years, several approaches have been proposed to stabilize high-order approximations.

In this internship, we focused on *Finite Volume Subcell* correction, introduced by François Vilar. The primary aim of this correction method is to maintain the high accuracy and precise subcell resolution of dG schemes. Therefore, an a posteriori correction will be used only when necessary at the subcell scale, while ensuring the conservation of the scheme. To achieve this, the dG scheme will be reformulated as a subcell FV scheme using the correct numerical flux, resulting in the dG reconstructed flux. This forms the basis of the limiter framework.

At each time step, a candidate solution is computed, and if it meets certain criteria (such as being positive and non-oscillating), the solution is accepted and the computation continues. If the solution is not admissible, the previous time step is returned to, and a local correction is made at the subcell scale. This is called a posteriori limitation. Each cell is divided into subcells, and if a subcell's solution is detected as problematic, a robust first-order or second-order TVD numerical flux is used on the subcell boundaries. If the subcell solution is admissible, the high-order reconstructed flux is used, retaining the dG scheme's accurate resolution and conservation properties. Only the

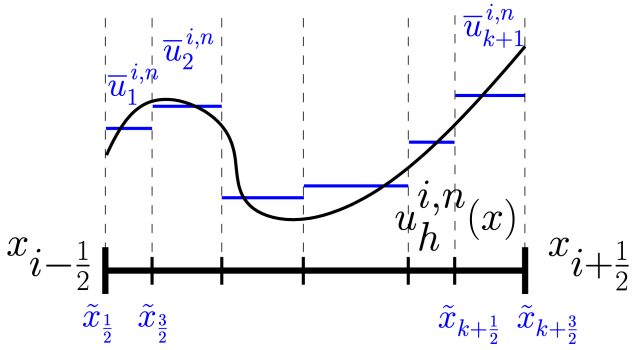


Figure 1.3: Example of subcell mean values on mesh $\omega_i := [x_{1-\frac{1}{2}}, x_{1+\frac{1}{2}}]$.

solution inside troubled subcells and its first neighbors are recomputed, while the rest of the solution remains unchanged. This a posteriori approach has been used by Ali Haidar in his PhD to deal with NSW equations interactions between surface waves and a floating structure. In this internship we will consider the a priori approach instead, developped later in this document.

1.4 Arbitrary Lagrangian-Eulerian description.

The *Arbitrary Lagrangian Eulerian* (ALE) model is a mathematical framework for modeling fluid dynamics and other physical systems. It is a hybrid approach to fluid dynamics that combines the Lagrangian and Eulerian approaches. The ALE method describes fluid motion with respect to a moving or deforming grid, which is useful for modeling fluid flows in complex geometries.

The Lagrangian approach to fluid dynamics is based on the idea of tracking individual fluid particles through space and time. The Eulerian approach, on the other hand, is based on describing the properties of the fluid at fixed points in space and time. Both approaches have advantages and disadvantages, and the ALE framework combines the best of both worlds.

Indeed, fluid motion is described in terms of a moving grid, which is typically defined by a set of control points. Based on the motion of the fluid, the position of these control points is updated at each time step. This allows the grid to move and deform along with the fluid, which is especially useful when modeling fluid flows in complex geometries like those found in industrial applications.

The ALE method allows for the grid to move and deform with the fluid, which is particularly useful in modeling shallow-water flows over complex geometries such as coastal regions, riverbeds, and estuaries. It can adapt the mesh to the flow gradients, which allows for a more efficient use of computational resources and accurate modeling of the flow behavior.

Section 2

Mathematical framework.

This section will introduce and define many mathematical objects we will use in the following work, especially notions related to discretization.

2.1 Discrete settings for dG method.

2.1.1 Spatial settings and basis functions.

Let $\Omega \in \mathbb{R}^d$, denote the computational domain of dimension d, with boundary $\partial\Omega$. In our work, we will only deal with the cases $d = 1$ and $d = 2$. We consider the following partition $\mathscr{T}_h := \{\omega_1, \ldots, \omega_{n_{\text{el}}}\} \subset \Omega$ of open disjoint elements ω of boundary $\partial\omega$ such that

$$\overline{\Omega} = \bigcup_{\omega \in \mathscr{T}_h} \overline{\omega}.$$

This partition is characterized by the mesh size $h := \max_{\omega \in \mathscr{T}_h} h_\omega$ where h_ω is the diameter of element ω. The dG methods are based on the following broken polynomial spaces, for an integer polynomial degree $k \geq 1$

$$\mathbb{P}^k(\mathscr{T}_h) := \{v \in L^2(\Omega) \mid \forall \omega \in \mathscr{T}_h,\ v_{|\omega} \in \mathbb{P}^k(\omega)\},$$

where $\mathbb{P}^k(\omega)$ denotes the space of polynomials in ω of total degree at most k, with $\dim(\mathbb{P}^k(\omega)) = k+1$. Piecewise polynomial functions belonging to $\mathbb{P}^k(\mathscr{T}_h)$ will be denoted with a subscript h in the following, and for all $\omega \in \mathscr{T}_h$ and $v_h \in \mathbb{P}^k(\mathscr{T}_h)$, we may use the shortcut $v_h^\omega := v_{h|\omega}$, when no confusion is possible.
For any mesh element $\omega \in \mathscr{T}_h$, we will consider a basis for $\mathbb{P}^k(\omega)$ denoted by

$$\mathbf{\Psi}_\omega := \{\psi_j^\omega\}_{j \in [\![1, k+1]\!]},$$

while the basis for global space $\mathbb{P}^k(\mathscr{T}_h)$ is obtained by taking the cartesian product[1] of the basis for the local polynomial spaces, i.e.

$$\mathbf{\Psi}_h := \bigtimes_{\omega\in\mathscr{T}_h} \mathbf{\Psi}_\omega = \left\{\{\psi_j^\omega\}_{j\in[\![1,k+1]\!]}\right\}_{\omega\in\mathscr{T}_h}.$$

We may also use the following shortcut notations for all $\omega_i \in \mathscr{T}_h$, and for smooth enough scalar-valued functions v and w :

$$\int_{\mathscr{T}_h} v(x)w(x)\,\mathrm{d}x := \sum_{\omega\in\mathscr{T}_h}\int_\omega v(x)w(x)\,\mathrm{d}x,$$
$$[v]_{\partial\omega_i} := v(x_{i+\frac{1}{2}}) - v(x_{i-\frac{1}{2}}).$$

For $\omega \in \mathscr{T}_h$, we denote by p_ω^k the L^2-orthogonal projector onto $\mathbb{P}^k(\omega)$ and $p_{\mathscr{T}_h}^k$ the global L^2-orthogonal projector onto $\mathbb{P}^k(\mathscr{T}_h)$ that gather all the local L^2 projectors p_ω^k on each element.

2.1.2 Time integration.

Let us also introduce a general partition of the time domain $0 = t^0 < t^1 < \cdots < t^n < \cdots < t^N = T$, and the time step $\Delta t^n = t^{n+1} - t^n$. Time integration may be carried out using explicit *Strong Stability Preserving Runge-Kutta* schemes. For instance, considering the following semi-discrete equation in operator form

$$\partial_t u_h + \mathscr{L}_h(u_h) = 0,$$

we advance from time level n to $(n+1)$ with the third-order scheme as follows :

$$\begin{aligned}
u_h^{n,1} &= u_h^n - \Delta t^n \mathscr{L}_h(u_h^n),\\
u_h^{n,2} &= \frac{1}{4}(3u_h^n + u_h^{n,1}) - \frac{\Delta t^n}{4}\mathscr{L}_h(u_h^{n,1}),\\
u_h^{n+1} &= \frac{1}{3}(u_h^n + 2u_h^{n,2}) - \frac{\Delta t^n}{3}\mathscr{L}_h(u_h^{n,2}),
\end{aligned}$$

where $u_h^{n,1}$ and $u_h^{n,2}$ are the solution obtained at intermediate stages, and the initial data u_h^0 is defined as the L^2 projection of the initial datum. Everytime Δt will be chosen according a CFL condition that will vary depending on the problem we will deal with. Typically, for dG methods with a Runge-Kutta scheme, our CFL condition could look like $\mathrm{CFL} = (2k-1)^{-1}$, k being the order of the method.

[1] Note that we have also, for all $\omega \in \mathscr{T}_h$ and $j \in [\![1, k+1]\!]$, $\mathrm{supp}(\psi_j^\omega) \subset \overline{\Omega}$.

2.2 Discontinuous Galerkin formulation for the NSW equations.

Let us remind the Nonlinear Shallow-Water equations :

$$\partial_t \mathbf{v} + \partial_x \mathbf{F}(\mathbf{v}, b) = \mathbf{B}(\mathbf{v}, \partial_x b), \tag{2.1}$$

where $\mathbf{v} : \mathbb{R} \times \mathbb{R}_+ \to \Theta$ gathers the flow's conservative variables and is assumed to take values in the convex and open set

$$\Theta = \{(\eta, q) \in \mathbb{R}^2 \mid H := \eta - b \geq 0\}, \tag{2.2}$$

where $\mathbf{F} : \Theta \times \mathbb{R} \to \mathbb{R}^2$ is the nonlinear flux function and $\mathbf{B} : \Theta \times \mathbb{R} \to \mathbb{R}^2$ is the topography source term, defined as follows

$$\mathbf{v} = (\eta, q)^T, \qquad \mathbf{F}(\mathbf{v}, b) = \left(q, uq + \frac{g(\eta^2 - 2\eta b)}{2}\right)^T, \qquad \mathbf{B}(\mathbf{v}, \partial_x b) = (0, -g\eta\partial_x b)^T.$$

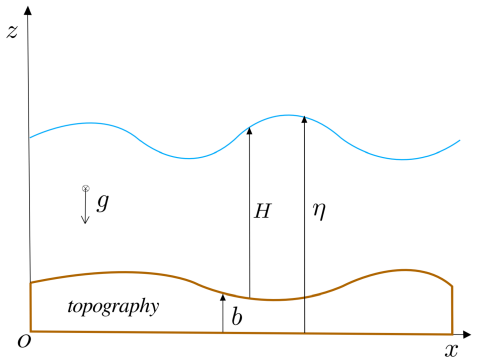


Figure 2.1: Main notations for free surface flows.

Let us now multiply (2.1) by a test function $\psi \in \mathbb{P}^k(\mathcal{T}_h)$, and integrate locally on cell ω_i to get

$$\int_{\omega_i} \psi \partial_t \mathbf{v} \, \mathrm{d}x + \int_{\omega_i} \psi \partial_x \mathbf{F}(\mathbf{v}, b) \, \mathrm{d}x = \int_{\omega_i} \psi \mathbf{B}(\mathbf{v}, \partial_x b) \, \mathrm{d}x,$$

and since the arbitrary function ψ is not time dependant we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\omega_i} \psi \mathbf{v} \, \mathrm{d}x + \int_{\omega_i} \psi \partial_x \mathbf{F}(\mathbf{v}, b) \, \mathrm{d}x = \int_{\omega_i} \psi \mathbf{B}(\mathbf{v}, \partial_x b) \, \mathrm{d}x.$$

An integration by part will then give us

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\omega_i} \psi \mathbf{v} \, \mathrm{d}x - \int_{\omega_i} \mathbf{F}(\mathbf{v}, b) \partial_x \psi \, \mathrm{d}x + \left[\psi \mathbf{F}(\mathbf{v}, b)\right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} = \int_{\omega_i} \psi \mathbf{B}(\mathbf{v}, \partial_x b) \, \mathrm{d}x.$$

Let us now consider the restrictions of the sought solution and the bathymetry to the mesh element

ω_i, which writes

$$\mathbf{v}_h^{\omega_i} = \sum_{p=1}^{k+1} \underline{\mathbf{v}}_p^{\omega_i} \psi_p^{\omega_i} \qquad \text{and} \qquad b_h^{\omega_i} = \sum_{p=1}^{k+1} \underline{b}_p^{\omega_i} \psi_p^{\omega_i}.$$

According to those, the dG local formulation can be written, for any $l \in [\![1, k+1]\!]$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\omega_i} \psi_l^{\omega_i} \mathbf{v}_h^{\omega_i} dx - \int_{\omega_i} \mathbf{F}(\mathbf{v}_h^{\omega_i}, b_h^{\omega_i}) \partial_x \psi_l^{\omega_i} \, \mathrm{d}x + \left[\psi_l^{\omega_i} \mathscr{F}\right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} = \int_{\omega_i} \psi_l^{\omega_i} \mathbf{B}(\mathbf{v}_h^{\omega_i}, \partial_x b_h^{\omega_i}) \, \mathrm{d}x,$$

where $\mathscr{F}$ is the numerical flux that needs to be consistent with the flux function $\mathbf{F}$. In the context of dG schemes, the numerical flux is defined as a function of the left and right traces of the local polynomial approximation coming from each side of the interface, i.e.

$$\mathscr{F}_{i+\frac{1}{2}} = \mathscr{F}\left(\mathbf{v}_h^{\omega_i}(x_{i+\frac{1}{2}}, t), \mathbf{v}_h^{\omega_{i+1}}(x_{i+\frac{1}{2}}, t)\right).$$

Here we will use the *Global Lax-Friedrichs* numerical flux which reads

$$\mathscr{F}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{F}(\mathbf{u}) + \mathbf{F}(\mathbf{v})}{2} - \frac{\sigma}{2}(\mathbf{v} - \mathbf{u}),$$

where $\sigma = \max_{\omega \in \mathscr{T}_h}(|u| + \sqrt{gH})$. Using the notations $\mathbf{v}_{i-\frac{1}{2}}^{\pm}$, $\mathbf{v}_{i+\frac{1}{2}}^{\pm}$ for the left and right traces of v_h in the boundaries $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ of ω_i, the previous formulation becomes, for any $l \in [\![1, k+1]\!]$:

$$\sum_{p=1}^{k+1} \partial_t \underline{\mathbf{v}}_p^{\omega_i} \int_{\omega_i} \psi_p^{\omega_i} \psi_l^{\omega_i} dx - \int_{\omega_i} \mathbf{F}(\mathbf{v}_h^{\omega_i}, b_h^{\omega_i}) \partial_x \psi_l^{\omega_i} \, \mathrm{d}x + \left[\psi_l^{\omega_i} \mathscr{F}\right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} = \int_{\omega_i} \psi_l^{\omega_i} \mathbf{B}(\mathbf{v}_h^{\omega_i}, \partial_x b_h^{\omega_i}) \, \mathrm{d}x, \tag{2.3}$$

and the global dG solution will be obtained by the gathering all the local solutions.

Remark 2.2.1. The local semi-discrete system (2.3) can be expressed in matrix form :

$$\mathbf{M}_{\omega_i} \partial_t \underline{v}_{\omega_i} = \widetilde{\mathbf{L}}_{\omega_i},$$

where $\mathbf{M}_{\omega_i}$ is the local *mass matrix*, i.e.

$$\mathbf{M}_{\omega_i} = \left(\mathbf{m}_{l,p}^{\omega_i}\right)_{l \in [\![1,k+1]\!]} = \left(\int_{\omega_i} \psi_l^{\omega_i} \psi_p^{\omega_i} \, \mathrm{d}x\right)_{l \in [\![1,k+1]\!]},$$

and $\widetilde{\mathbf{L}}_{\omega_i}$ being the local residual vector gathering the volume integrals, the surface integrals and source terms as follows :

$$\widetilde{\mathbf{L}}_{\omega_i} = \left(\boldsymbol{\ell}_{l,p}^{\omega_i}\right)_{l \in [\![1,k+1]\!]} = \left(\int_{\omega_i} \mathbf{F}(\mathbf{v}_h^{\omega_i}, b_h^{\omega_i}) \partial_x \psi_l^{\omega_i} \, \mathrm{d}x - \left[\psi_l^{\omega_i} \mathscr{F}\right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} + \int_{\omega_i} \psi_l^{\omega_i} \mathbf{B}(\mathbf{v}_h^{\omega_i}, \partial_x b_h^{\omega_i}) \, \mathrm{d}x\right)_{l \in [\![1,k+1]\!]}$$

Remark 2.2.2. Same way as previously, we can express the global semi-discrete system in matrix form, i.e.

$$\mathbf{M} \partial_t \underline{v} = \widetilde{\mathbf{L}},$$

where $\mathbf{M}$ is the block-diagonal matrix defined like this

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\omega_1} & 0 & \dots & 0 \\ 0 & \mathbf{M}_{\omega_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{M}_{\omega_{n_{\text{el}}}} \end{bmatrix} = \begin{bmatrix} \left(\mathbf{m}_{l,p}^{\omega_1}\right)_{l\in[\![1,k+1]\!]} & 0 & \dots & 0 \\ 0 & \left(\mathbf{m}_{l,p}^{\omega_2}\right)_{l\in[\![1,k+1]\!]} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \left(\mathbf{m}_{l,p}^{\omega_{n_{\text{el}}}}\right)_{l\in[\![1,k+1]\!]} \end{bmatrix},$$

and

$$\widetilde{\mathbf{L}} = \left[\widetilde{\mathbf{L}}_{\omega_1}\, \widetilde{\mathbf{L}}_{\omega_2} \dots \widetilde{\mathbf{L}}_{\omega_{n_{\text{el}}}}\right]^T = \left[\left(\boldsymbol{\ell}_{l,p}^{\omega_1}\right)_{l\in[\![1,k+1]\!]} \left(\boldsymbol{\ell}_{l,p}^{\omega_2}\right)_{l\in[\![1,k+1]\!]} \dots \left(\boldsymbol{\ell}_{l,p}^{\omega_{n_{\text{el}}}}\right)_{l\in[\![1,k+1]\!]}\right]^T.$$

2.3 Coupling with ALE description.

Here we introduce an ALE description that will be used to deal with the coupled problem for the Ph.D. A central aspect of any ALE description is the construction of a continuous and regular coordinate transformation, allowing to recast the equations from the initial (stationary) domain Ω_0 to the current moving domain Ω_t :

$$\Omega_0 \times [0, T_{\max}] \ni (\boldsymbol{X}, t) \mapsto \boldsymbol{x}(\boldsymbol{X}, t) \in \Omega_t,$$

where $\boldsymbol{X}$ refers to the *reference* coordinate in Ω_0 and $\boldsymbol{x} := \boldsymbol{x}(\boldsymbol{X}, t)$ the associated *physical* coordinate in Ω_t. Further assuming this mapping to be continuously differentiable with respect to time, piecewise continuously differentiable with respect to $\boldsymbol{X}$, and denoting by $\mathsf{v}_g(\boldsymbol{x}, t)$ the grid's velocity at the physical point $\boldsymbol{x} := \boldsymbol{x}(\boldsymbol{X}, t)$, we have the following identity :

$$\mathsf{v}_g(\boldsymbol{x}(\boldsymbol{X}, t), t) = \partial_t \boldsymbol{x}(\boldsymbol{X}, t).$$

Now, for the sake of notations, considering any function $v(\boldsymbol{x}, t)$, we introduce $\widetilde{v}(\boldsymbol{X}, t)$ its counterpart defined on the referential frame as

$$\widetilde{v}(\boldsymbol{X}, t) := v(\boldsymbol{x}(\boldsymbol{X}, t), t).$$

Then, for any arbitrary and regular enough function $v(\boldsymbol{x}, t)$, the fundamental ALE relation between the total time derivative, the Eulerian time derivative and the spatial derivative is

$$\frac{\mathrm{d}}{\mathrm{d}t} v(\boldsymbol{x}(\boldsymbol{X}, t), t) := \partial_t v(\boldsymbol{x}(\boldsymbol{X}, t), t) + \mathsf{v}_g \partial_x v(\boldsymbol{x}(\boldsymbol{X}, t), t) = (\partial_t + \mathsf{v}_g \partial_x) v(\boldsymbol{x}(\boldsymbol{X}, t), t) =: \partial_t \widetilde{v}(\boldsymbol{X}, t).$$

Section 3

Finite-volume local subcell correction.

3.1 Subcell formulation on scalar conservation law.

More details about this section can be found in [4]. Some mathematicians discovered a remarkable equivalence of general diagonal norm high-order summation-by-parts operators to a finite volume formulation based on subcells. By using a specific set of subcell finite volume fluxes, they can construct provably entropy stable schemes. It also shows scheme conservation at the subcell level. This subcell finite volume formulation is notable in that it allows one to directly influence scheme properties such as entropy stability by selecting the appropriate subcell finite volume fluxes. Now Let us consider a subcell finite volume formulation for general dG schemes. Here we take the scalar conservation law

$$\begin{aligned} &\partial_t u(t,x) + \partial_x f(u(t,x)) = 0, \qquad t \in [0,T], \;\; x \in \Omega. \\ &u(0,x) = u_0(x), x \in \Omega. \end{aligned}$$

Let $\varphi \in \mathscr{C}_c^\infty(\Omega)$ a test function. Multiplying by φ and integrating the SCL over ω_i gives us

$$\int_{\omega_i} \varphi \partial_t u \, \mathrm{d}x + \int_{\omega_i} \partial_x f(u) \varphi \, \mathrm{d}x = 0,$$

and integrating by parts leads to

$$\int_{\omega_i} \varphi \partial_t u \, \mathrm{d}x = \int_{\omega_i} f(u) \partial_x \varphi \, \mathrm{d}x - \left[f(u)\varphi \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}.$$

Let us now divide our domain into subcells, according to (2.1), such that for any mesh element $\omega_i \in \mathscr{T}_h$, we introduce a sub-partition $\mathscr{T}_{\omega_i}$ into $k+1$ open disjoint subcells :

$$\overline{\omega}_i := \bigcup_{m=1}^{k+1} \overline{S}_m^{\omega_i} = \bigcup_{m=1}^{k+1} \left[\widetilde{x}_{i-\frac{1}{2}}^{\omega_i}, \widetilde{x}_{i+\frac{1}{2}}^{\omega_i} \right]$$

where the subcell $S_m^{\omega_i} := \left[\widetilde{x}_{i-\frac{1}{2}}^{\omega_i}, \widetilde{x}_{i+\frac{1}{2}}^{\omega_i}\right]$ is of size $|S_m^{\omega_i}| = \left|\widetilde{x}_{i+\frac{1}{2}}^{\omega_i} - \widetilde{x}_{i-\frac{1}{2}}^{\omega_i}\right|$, with the convention $\widetilde{x}_{\frac{1}{2}}^{\omega_i} = x_{i-\frac{1}{2}}$ and $\widetilde{x}_{k+\frac{3}{2}}^{\omega_i} = x_{i+\frac{1}{2}}$. When considering a sequence of neighboring element ω_{i-1}, ω_i and ω_{i+1}, we may use the convenient convention $S_0^{\omega_i} := S_{k+1}^{\omega_{i-1}}$ and $S_{k+2}^{\omega_i} := S_1^{\omega_{i+1}}$.

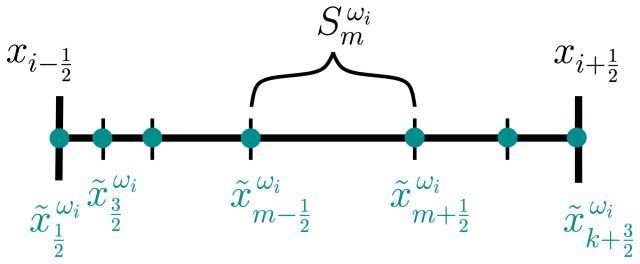


Figure 3.1: Subcell decomposition of ω_i through $k+2$ flux points.

Remark 3.1.1. Let us keep in mind that these flux points can be chosen completely arbitrarily and are not required to be related to any quadrature rule or diagonal norm matrix, as is typically the case in entropy stable schemes. Furthermore, the choice of the dG basis functions has no effect on these flux points. Let us reiterate that any cell subdivision would result in the same theoretical result, namely the same dG scheme. It would only have an impact on the definition of the subcell finite volume numerical fluxes defined in the rest of the current development. However, it has an effect on the correction procedure, as we'll see later.

Now, the important step here is the introduction of some very specific basis functions that we refer from now on to as *subresolution basis functions*. For an $\omega \in \mathscr{T}_h$, these functions can be seen as the L^2 projection onto $\mathbb{P}^k(\omega)$ of the subcell indicator function $\{\mathbf{1}_m^\omega \mid m \in [\![1, k+1]\!]\}$ such that $\mathbf{1}_m^\omega(x) = 1$ if $x \in S_i^\omega$, and 0 otherwise. Then we have the set of subresolution basis functions $\{\phi_m^\omega \in \mathbb{P}^k(\omega) \mid m \in [\![1, k+1]\!]\}$ where the basis is defined as follows

$$\phi_m^\omega = p_\omega^k(\mathbf{1}_m^\omega), \qquad m \in [\![1, k+1]\!], \tag{3.1}$$

such that for any $\varphi \in \mathbb{P}^k(\omega)$, we have

$$\int_\omega \phi_m^\omega \varphi \, \mathrm{d}x = \int_\omega \mathbf{1}_m^\omega \varphi \, \mathrm{d}x = \int_{S_i^\omega} \varphi \, \mathrm{d}x.$$

Let us also denote that this condition enforces that for all $x \in \omega_i$,

$$\sum_{m=1}^{k+1} \phi_m^{\omega_i}(x) = 1.$$

Finally, for every $\omega \in \mathscr{T}_h$, we introduce the set of piecewise constant functions on the sub-grid :

$$\mathbb{P}^0(\mathscr{T}_h) := \{v \in L^2(\omega) \ \mid \forall S_m^\omega \in \mathscr{T}_\omega, \ v_{|S_m^\omega} \in \mathbb{P}^0(S_m^\omega)\}.$$

Here we will consider the classical time discretization and marching algorithms given in (2.1.2). Now, in order to obtain a $(k+1)^{th}$ order discretization, let us consider here a piecewise polynomial approximated solution $u_h(x,t)$, where its restriction to cell ω_i, namely $u_{h|\omega_i} = u_h^{\omega_i}$, belongs to $\mathbb{P}^k(\omega_i)$. Then the solution writes

$$u_h^{\omega_i}(x,t) = \sum_{m=1}^{k+1} u_m^{\omega_i}(t)\phi_m^{\omega_i}(t),$$

the coefficients $u_m^{\omega_i}$ being the solution moments to be computed through a local variational formulation on ω_i, i.e. find $u_h^{\omega_i} \in \mathbb{P}^k(\omega_i)$ such that, for all $\varphi \in \mathbb{P}^k(\omega_i)$,

$$\int_{\omega_i} \varphi \partial_t u_h^{\omega_i} \, \mathrm{d}x = \int_{\omega_i} f(u_h^{\omega_i}) \partial_x \varphi \, \mathrm{d}x - \left[\mathscr{F}\varphi\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}. \tag{3.2}$$

where $\mathscr{F}$ is the dG numerical flux function[1]. In this context, $\mathscr{F}$ is defined as a function of the two states on the left and right of each interface, i.e. $\mathscr{F}_{i-\frac{1}{2}} = \mathscr{F}\left(u_h^{\omega_i}(x_{i+\frac{1}{2}},t), u_h^{\omega_{i+1}}(x_{i+\frac{1}{2}},t)\right)$. This function is generally obtained by solving an exact or approximated *Riemann problem*. In this section we will only consider the classical *local Lax-Friedrichs* numerical flux, which reads

$$\mathscr{F}(u,v) = \frac{f(u)+f(v)}{2} - \frac{\gamma(u,v)}{2}(v-u),$$

where $\gamma(u,v) = \max(|f'(u)|, |f'(v)|)$.

Here in (3.2), we need to substitute in the volume integral $\int_{\omega_i} f(u_h^{\omega_i})\partial_x\varphi \, \mathrm{d}x$ the exact interior flux function $f(u_h^{\omega_i})$ with some polynomial approximation $f_h^{\omega_i}$. To achieve this, we define $f_h^{\omega_i} \in \mathbb{P}^\alpha(\omega_i)$, where $\alpha \in \mathbb{N}^*$, as the L^2 projection of function $f(u_h^{\omega_i})$ onto $\mathbb{P}^\alpha(\omega_i)$ as follows

$$\int_{\omega_i} f_h^{\omega_i}\psi \, \mathrm{d}x = \int_{\omega_i} f(u_h^{\omega_i})\psi \, \mathrm{d}x, \qquad \forall \psi \in \mathbb{P}^\alpha(\omega_i).$$

We need that $\alpha \geq k-1$, and that $\int_{\omega_i} f(u_h^{\omega_i})\psi \, \mathrm{d}x$ is computed the same way[2] as the volume integral $\int_{\omega_i} f(u_h^{\omega_i})\partial_x\varphi \, \mathrm{d}x$ in (3.2), to ensure that the dG schemes can rewrites as

$$\int_{\omega_i} \varphi \partial_t u_h^{\omega_i} \, \mathrm{d}x = \int_{\omega_i} f_h^{\omega_i}\partial_x\varphi \, \mathrm{d}x - \left[\mathscr{F}\varphi\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}, \qquad \forall \varphi \in \mathbb{P}^k(\omega_i).$$

Therefore, through analytical integration or using quadrature, the volume integral $\int_{\omega_i} f_h^{\omega_i}\partial_x\varphi \, \mathrm{d}x$ will be computed exactly, as $f_h^{\omega_i}\partial_x\varphi \, \mathrm{d}x \in \mathbb{P}^{\alpha+k-1}(\omega_i)$. We get then, for all $\varphi \in \mathbb{P}^k(\omega_i)$, the following strong form

$$\int_{\omega_i} \varphi \partial_t u_h^{\omega_i} \, \mathrm{d}x = \int_{\omega_i} f_h^{\omega_i}\partial_x\varphi \, \mathrm{d}x - \left[\mathscr{F}\varphi\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = -\int_{\omega_i} \partial_x f_h^{\omega_i}\varphi \, \mathrm{d}x + \left[(f_h^i - \mathscr{F})\varphi\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}. \tag{3.3}$$

[1] In addition to ensure the scheme conservation, the numerical flux is the cornerstone of any Fv or dG scheme regarding fundamental considerations as stability, positivity and entropy among others.

[2] Namely by exact integration, or by the same quadrature rule. In dG schemes, volume integrals are generally computed by means of a quadrature rule. And for the purpose of accuracy, it is possible to demonstrate that to design a $(k+1)^{th}$ order numerical scheme, a quadrature rule exact at least for polynomial up to degree $2k$ is required.

Since the subresolution basis function are L^2 projections and equation (3.3) holds for any polynomial of degree k, Let us replace φ by $\phi_m^{\omega_i}$ and get

$$\int_{\omega_i} \phi_m^{\omega_i} \partial_t u_h^{\omega_i} \, \mathrm{d}x = -\int_{\omega_i} \partial_x f_h^{\omega_i} \phi_m^{\omega_i} \, \mathrm{d}x + \left[(f_h^i - \mathscr{F}) \phi_m^{\omega_i} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}, \qquad m \in [\![1, k+1]\!].$$

Recalling that $\partial_t u_h^{\omega_i}, \partial_x f_h^{\omega_i} \in \mathbb{P}^k(\omega_i)$, we have

$$\int_{\omega_i} \phi_m^{\omega_i} \partial_t u_h^{\omega_i} \, \mathrm{d}x = \int_{S_m^{\omega_i}} \partial_t u_h^{\omega_i} \, \mathrm{d}x \qquad \text{and} \qquad \int_{\omega_i} \partial_x f_h^{\omega_i} \phi_m^{\omega_i} \, \mathrm{d}x = \int_{S_m^{\omega_i}} \partial_x f_h^{\omega_i} \, \mathrm{d}x,$$

and introducing the *subcell mean values* $\overline{u}_m^{\omega_i}$ such that

$$\overline{u}_m^{\omega_i} = \frac{1}{|S_m^{\omega_i}|} \int_{S_m^{\omega_i}} u_h^{\omega_i} \, \mathrm{d}x,$$

we can get

$$\begin{aligned} &\int_{\omega_i} \phi_m^{\omega_i} \partial_t u_h^{\omega_i} \, \mathrm{d}x = -\int_{\omega_i} \partial_x f_h^{\omega_i} \phi_m^{\omega_i} \, \mathrm{d}x + \left[(f_h^{\omega_i} - \mathscr{F}) \phi_m^{\omega_i} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \\ \Leftrightarrow &\int_{S_m^{\omega_i}} \partial_t u_h^{\omega_i} \, \mathrm{d}x = -\int_{S_m^{\omega_i}} \partial_x f_h^{\omega_i} \, \mathrm{d}x + \left[(f_h^{\omega_i} - \mathscr{F}) \phi_m^{\omega_i} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}, \end{aligned}$$

and finally obtain the scheme

$$\partial_t \overline{u}_m^{\omega_i} = -\frac{1}{|S_m^{\omega_i}|} \left(\left[f_h^{\omega_i} \right]_{\widetilde{x}_{m-\frac{1}{2}}}^{\widetilde{x}_{m+\frac{1}{2}}} - \left[(f_h^{\omega_i} - \mathscr{F}) \phi_m^{\omega_i} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \right).$$

The last step is to introduce the $k+2$ subcell *finite volume like* flux $\left\{ \widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} \right\}_{m \in [\![0,k+1]\!]}$ that will be referred now as the *reconstructed fluxes*, located at the $k+2$ flux points. These reconstructed fluxes are defined through the following linear system :

$$\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - \widehat{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i} = \left[f_h^i \right]_{\widetilde{x}_{m-\frac{1}{2}}}^{\widetilde{x}_{m+\frac{1}{2}}} - \left[(f_h^i - \mathscr{F}) \phi_m^{\omega_i} \right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}, \qquad m \in [\![0, k+1]\!] \tag{3.4}$$

$$\widehat{\mathscr{F}}_{\frac{1}{2}}^{\omega_i} = \mathscr{F}_{i-\frac{1}{2}}, \qquad \widehat{\mathscr{F}}_{k+\frac{3}{2}}^{\omega_i} = \mathscr{F}_{i+\frac{1}{2}}. \tag{3.5}$$

Solving this is pretty straightforward, as if we substitute m by p in (3.4) and summing on p from 1 to m leads to

$$\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} = f_h^{\omega_i}(\widetilde{x}_{m+\frac{1}{2}}^{\omega_i}) - \left(1 - \sum_{p=1}^{m} \phi_p^{\omega_i}(x_{i-\frac{1}{2}}) \right) \left(f_h^{\omega_i}(x_{i-\frac{1}{2}}) - \mathscr{F}_{i-\frac{1}{2}} \right) - \left(\sum_{p=1}^{m} \phi_p^{\omega_i}(x_{i+\frac{1}{2}}) \right) \left(f_h^{\omega_i}(x_{i+\frac{1}{2}}) - \mathscr{F}_{i+\frac{1}{2}} \right).$$

We finally get the value of the reconstructed fluxes :

$$\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} = f_h^{\omega_i}(\widetilde{x}_{m+\frac{1}{2}}^{\omega_i}) - C_{m+\frac{1}{2}}^{i-\frac{1}{2}} \left(f_h^{\omega_i}(x_{i-\frac{1}{2}}) - \mathscr{F}_{i-\frac{1}{2}} \right) - C_{m+\frac{1}{2}}^{i+\frac{1}{2}} \left(f_h^{\omega_i}(x_{i+\frac{1}{2}}) - \mathscr{F}_{i+\frac{1}{2}} \right), \tag{3.6}$$

where the correction coefficients $C_{m+\frac{1}{2}}^{i\pm\frac{1}{2}}$ are defined as

$$C_{m+\frac{1}{2}}^{i-\frac{1}{2}} := \sum_{p=m+1}^{k+1} \phi_p^{\omega_i}(x_{i-\frac{1}{2}}) \qquad \text{and} \qquad C_{m+\frac{1}{2}}^{i-\frac{1}{2}} := \sum_{p=1}^{m} \phi_p^{\omega_i}(x_{i+\frac{1}{2}}).$$

Those reconstructed flux (3.6) are nothing but the interior polynomial flux $f_h^{\omega_i}(\widetilde{x}_{m+\frac{1}{2}}^{\omega_i})$ with some corrections terms that take account of the difference between the boundary values of this interior flux and the numerical flux.
Let us finally gather all of these in the following result :

Theorem 3.1.2. *Provided the analytical calculation of volume integrals, or alternatively by means of quadrature rule, discontinuous Galerkin schemes expressed in cell ω_i as follows*

$$\int_{\omega_i} \varphi \partial_t u_h^{\omega_i} \, \mathrm{d}x = \int_{\omega_i} f(u_h^{\omega_i}) \partial_x \varphi \, \mathrm{d}x - \left[\mathscr{F}\varphi\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}, \qquad \forall \varphi \in \mathbb{P}^k(\omega_i),$$

can be recast into $k+1$ subcell finite volume schemes as, for $m \in [\![1, k+1]\!]$

$$\partial_t \overline{u}_m^{\omega_i} = -\frac{1}{|S_m^{\omega_i}|} \left(\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - \widehat{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i} \right),$$

where the $k+2$ reconstructed fluxes $\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i}$ are defined by

$$\widehat{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} = f_h^{\omega_i}(\widetilde{x}_{m+\frac{1}{2}}^{\omega_i}) - C_{m+\frac{1}{2}}^{i-\frac{1}{2}} \left(f_h^{\omega_i}(x_{i-\frac{1}{2}}) - \mathscr{F}_{i-\frac{1}{2}} \right) - C_{m+\frac{1}{2}}^{i+\frac{1}{2}} \left(f_h^{\omega_i}(x_{i+\frac{1}{2}}) - \mathscr{F}_{i+\frac{1}{2}} \right).$$

In the last expression, for $\alpha \in [\![k-1, k+1]\!]$, the polynomial flux $f_h^{\omega_i}$ is either a L^2 projection of $f(u_h^{\omega_i})$ onto $\in \mathbb{P}^\alpha(\omega_i)$, or collocated at $\alpha+1$ given points, as it is the case for collocated and nodal dG schemes. Simple explicit expression of the correction coefficients can be found in [4].

3.2 *A posteriori* approach.

To summarize, *a posteriori correction* is only applied locally at the subcell level where it is absolutely needed (i.e. only non-admissible subcells are marked), while not neglecting the scheme conservation property.

In practice, we first reformulate dG scheme as a FV-like subcell schemes provided the use of the so-called dG reconstructed flux. Then, the correction procedure is done as follows : at each SSP-RK time-step, we compute a high-order dG candidate solution and check its admissibility (non-negative water-height and no spurious oscillations). If the solution is admissible, we go further in time. If it is not the case, we go back to the previous time-step and correct locally at the subcell level the non-admissible local numerical solution.

Actually, we divide the cell into subcells, then, if the solution at a specific subcell is detected as bad, we substitute the dG reconstructed flux on the subcell boundaries by a robust low-order

FV numerical flux. Otherwise, if the solution is detected as admissible on this subcell, we keep the high-order dG reconstructed numerical flux.

The purpose of applying this correction procedure is to enforce the water-height positivity in Shallow-Water case and to avoid spurious oscillations in the vicinity of solution's singularity while preserving as much as possible the high accuracy and the very precise subcell resolution of high-order dG schemes, by minimizing the number of subcells in which the solution has to be recomputed. To this end, we pay attention to keep the scheme local conservation property. Not only the solutions are recomputed in the troubled subcells, but also in its first neighbors, so that we can have the same numerical fluxes on both sides of subcells interfaces. Reader can refer to [4] for more informations and applications to NSW in [2, 3].

3.3 *A priori* approach.

The main goal of this internship is to consider the *a priori* approach instead, a monolithic subcell dG/FV convex property preserving scheme that allows us to obtain good results with only few cells, without any need of coming back in time loop as the other approach, and lead to a simpler implementation.

3.3.1 Introduction of blended fluxes.

Here we will still consider *high-order reconstructed fluxes* $\widehat{\mathscr{F}}^{\omega_i}_{m+\frac{1}{2}}$ and we introduce $F^*_{m+\frac{1}{2}} = F^*(\overline{u}^{\omega_i}_m, \overline{u}^{\omega_i}_{m+1})$ *first-order finite volume flux* (Lax-Friedrichs for example). The idea of this approach is to assemble those two flux together into a new *blended* flux $\widetilde{\mathscr{F}}$, such that we can get the following subcell scheme

$$\begin{aligned}\partial_t \overline{u}^{\omega_i}_m &= -\frac{1}{|S^{\omega_i}_m|}\left(F^*_{m+\frac{1}{2}} - F^*_{m-\frac{1}{2}} + \theta_{m+\frac{1}{2}}\left(\widehat{\mathscr{F}}^{\omega_i}_{m+\frac{1}{2}} - F^*_{m+\frac{1}{2}}\right) - \theta_{m-\frac{1}{2}}\left(\widehat{\mathscr{F}}^{\omega_i}_{m-\frac{1}{2}} - F^*_{m-\frac{1}{2}}\right)\right)\\ &= -\frac{1}{|S^{\omega_i}_m|}\left(\widetilde{\mathscr{F}}^{\omega_i}_{m+\frac{1}{2}} - \widetilde{\mathscr{F}}^{\omega_i}_{m-\frac{1}{2}}\right),\end{aligned}$$

where $\widetilde{\mathscr{F}}$ is defined as

$$\widetilde{\mathscr{F}}^{\omega_i}_{m\pm\frac{1}{2}} = F^*_{m\pm\frac{1}{2}} + \theta_{m\pm\frac{1}{2}}\left(\widehat{\mathscr{F}}^{\omega_i}_{m\pm\frac{1}{2}} - F^*_{m\pm\frac{1}{2}}\right).$$

This scheme is by construction conservative at subcell level. In the following, we may use the following shortcut for clarity

$$\Delta\mathscr{F}_{m\pm\frac{1}{2}} := \widehat{\mathscr{F}}^{\omega_i}_{m\pm\frac{1}{2}} - F^*_{m\pm\frac{1}{2}},$$

with

$$F^*_{m+\frac{1}{2}} = F^*(\overline{u}^{\omega_i}_m, \overline{u}^{\omega_i}_{m+1}) = \frac{f(\overline{u}^{\omega_i}_m) + f(\overline{u}^{\omega_i}_{m+1})}{2} - \frac{\gamma_{m+\frac{1}{2}}}{2}(\overline{u}^{\omega_i}_{m+1} - \overline{u}^{\omega_i}_m).$$

Considering the previous scheme and using Forward-Euler[3], we can get

$$\partial_t \overline{u}_m^{\omega_i} = -\frac{1}{|S_m^{\omega_i}|}\left(\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - \widetilde{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i}\right) \Rightarrow \overline{u}_m^{\omega_i,n+1} = \overline{u}_m^{\omega_i,n} - \frac{\Delta t}{|S_m^{\omega_i}|}\left(\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - \widetilde{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i}\right).$$

Now the trick here is to add then substract some terms in order to rearrange our formula, i.e.

$$\begin{aligned}
\overline{u}_m^{\omega_i,n+1} &= \overline{u}_m^{\omega_i,n} - \frac{\Delta t}{|S_m^{\omega_i}|}\left(\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - \widetilde{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i}\right) \pm f(\overline{u}_m^{\omega_i,n}) \pm \frac{\Delta t}{|S_m^{\omega_i}|}(\gamma_{m+\frac{1}{2}} - \gamma_{m-\frac{1}{2}})\overline{u}_m^{\omega_i,n} \\
&= \left(1 - \frac{\Delta t}{|S_m^{\omega_i}|}(\gamma_{m+\frac{1}{2}} - \gamma_{m-\frac{1}{2}})\right)\overline{u}_m^{\omega_i,n} + \frac{\Delta t}{|S_m^{\omega_i}|}\gamma_{m+\frac{1}{2}}\left(\overline{u}_m^{\omega_i,n} - \frac{\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - f(\overline{u}_m^{\omega_i,n})}{\gamma_{m+\frac{1}{2}}}\right) \\
&+ \frac{\Delta t}{|S_m^{\omega_i}|}\gamma_{m-\frac{1}{2}}\left(\overline{u}_m^{\omega_i,n} + \frac{\widetilde{\mathscr{F}}_{m-\frac{1}{2}}^{\omega_i} - f(\overline{u}_m^{\omega_i,n})}{\gamma_{m-\frac{1}{2}}}\right),
\end{aligned}$$

where we will now using the shortcuts

$$\widetilde{u}_{m+\frac{1}{2}}^{*,-} := \overline{u}_m^{\omega_i,n} - \frac{\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - f(\overline{u}_m^{\omega_i,n})}{\gamma_{m+\frac{1}{2}}}, \qquad \widetilde{u}_{m-\frac{1}{2}}^{*,+} := \overline{u}_m^{\omega_i,n} + \frac{\widetilde{\mathscr{F}}_{m+\frac{1}{2}}^{\omega_i} - f(\overline{u}_m^{\omega_i,n})}{\gamma_{m+\frac{1}{2}}},$$

giving us the following

$$\overline{u}_m^{\omega_i,n+1} = \left(1 - \frac{\Delta t}{|S_m^{\omega_i}|}(\gamma_{m+\frac{1}{2}} - \gamma_{m-\frac{1}{2}})\right)\overline{u}_m^{\omega_i,n} + \frac{\Delta t}{|S_m^{\omega_i}|}\gamma_{m+\frac{1}{2}}\widetilde{u}_{m+\frac{1}{2}}^{*,-} + \frac{\Delta t}{|S_m^{\omega_i}|}\gamma_{m-\frac{1}{2}}\widetilde{u}_{m-\frac{1}{2}}^{*,+}.$$

Then, under the following CFL condition :

$$\Delta t \leq \frac{|S_m^{\omega_i}|}{\gamma_{m+\frac{1}{2}} + \gamma_{m-\frac{1}{2}}},$$

we have that $\overline{u}_m^{\omega_i,n+1}$ is a convex combination of $\overline{u}_m^{\omega_i,n}$, $\widetilde{u}_{m-\frac{1}{2}}^{*,+}$ and $\widetilde{u}_{m+\frac{1}{2}}^{*,-}$.

3.3.2 Conservation of max principle.

Now we need to assure that for a certain Θ being an admissible states convex set, we have the following assertion :

$$\forall (i,n) \in [\![1, n_{\text{el}}]\!] \times [0,T], \quad (\overline{u}_m^{\omega_i,n} \in \Theta) \Rightarrow \left(\widetilde{u}_{m\pm\frac{1}{2}}^{*,\pm} \in \Theta\right).$$

[3]We focus on Forward-Euler (FE) time stepping, as SSP Runge-Kutta can be formulated as convex combinations of FE.

First, Let us remark that we can write

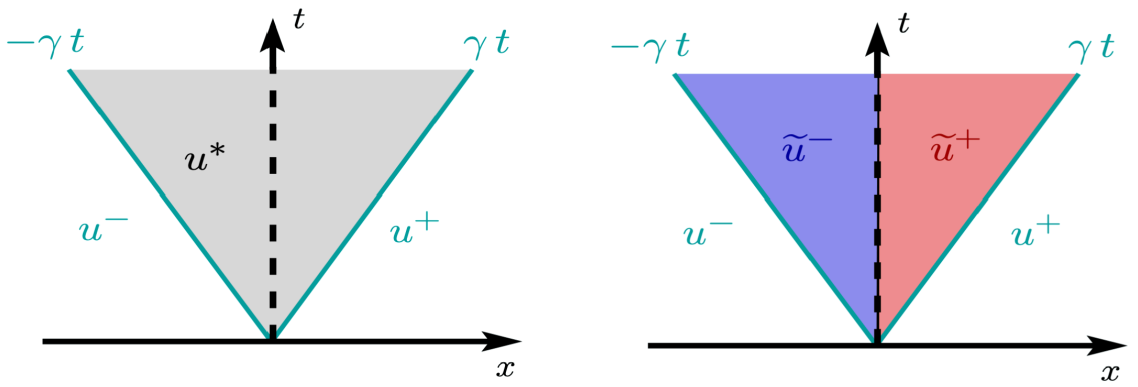
$$\begin{aligned}\widetilde{u}^{*,-}_{m+\frac{1}{2}} &= \overline{u}^{\omega_i,n}_m - \frac{\widetilde{\mathscr{F}}^{\omega_i}_{m+\frac{1}{2}} - f(\overline{u}^{\omega_i,n}_m)}{\gamma_{m+\frac{1}{2}}} = \overline{u}^{\omega_i,n}_m - \frac{F^*_{m+\frac{1}{2}} - f(\overline{u}^{\omega_i,n}_m)}{\gamma_{m+\frac{1}{2}}} - \theta_{m+\frac{1}{2}}\frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} \\ &= \frac{\overline{u}^{\omega_i,n}_m + \overline{u}^{\omega_i,n}_{m+1}}{2} - \frac{f(\overline{u}^{\omega_i,n}_{m+1}) - f(\overline{u}^{\omega_i,n}_m)}{2\gamma_{m+\frac{1}{2}}} - \theta_{m+\frac{1}{2}}\frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} \\ &= u^*_{m+\frac{1}{2}} - \theta_{m+\frac{1}{2}}\frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}},\end{aligned}$$

and

$$\begin{aligned}\widetilde{u}^{*,+}_{m-\frac{1}{2}} &= \overline{u}^{\omega_i,n}_m + \frac{\widetilde{\mathscr{F}}^{\omega_i}_{m-\frac{1}{2}} - f(\overline{u}^{\omega_i,n}_m)}{\gamma_{m-\frac{1}{2}}} = \overline{u}^{\omega_i,n}_m + \frac{F^*_{m-\frac{1}{2}} - f(\overline{u}^{\omega_i,n}_m)}{\gamma_{m-\frac{1}{2}}} + \theta_{m+\frac{1}{2}}\frac{\Delta\mathscr{F}_{m-\frac{1}{2}}}{\gamma_{m-\frac{1}{2}}} \\ &= \frac{\overline{u}^{\omega_i,n}_m + \overline{u}^{\omega_i,n}_{m-1}}{2} - \frac{f(\overline{u}^{\omega_i,n}_m) - f(\overline{u}^{\omega_i,n}_{m-1})}{2\gamma_{m-\frac{1}{2}}} + \theta_{m-\frac{1}{2}}\frac{\Delta\mathscr{F}_{m-\frac{1}{2}}}{\gamma_{m-\frac{1}{2}}} = u^*_{m-\frac{1}{2}} + \theta_{m-\frac{1}{2}}\frac{\Delta\mathscr{F}_{m-\frac{1}{2}}}{\gamma_{m-\frac{1}{2}}},\end{aligned}$$

with

$$\begin{aligned}u^*_{m+\frac{1}{2}} &:= \frac{\overline{u}^{\omega_i,n}_m + \overline{u}^{\omega_i,n}_{m+1}}{2} - \frac{f(\overline{u}^{\omega_i,n}_{m+1}) - f(\overline{u}^{\omega_i,n}_m)}{2\gamma_{m+\frac{1}{2}}} \\ u^*_{m-\frac{1}{2}} &:= \frac{\overline{u}^{\omega_i,n}_m + \overline{u}^{\omega_i,n}_{m-1}}{2} - \frac{f(\overline{u}^{\omega_i,n}_m) - f(\overline{u}^{\omega_i,n}_{m-1})}{2\gamma_{m-\frac{1}{2}}}.\end{aligned}$$



(a) 1^{st} order situation. (b) Blended flux situation.

Figure 3.2: Modified Riemann intermediate states.

This way, since FV being robust and conserving max principles, it is obvious that the Riemann intermediate states $u^*_{m+\frac{1}{2}}, u^*_{m-\frac{1}{2}} \in \Theta$. Here, we need to assure that $\widetilde{u}^{*,+}_{m-\frac{1}{2}}, \widetilde{u}^{*,-}_{m+\frac{1}{2}} \in [u^m_{\min}, u^m_{\max}] \subset [u_{\min}, u_{\max}]$. Then it implies that we should have

$$\widetilde{u}^{*,-}_{m+\frac{1}{2}} \in [u^m_{\min}, u^m_{\max}] \supset \mathscr{I}(\overline{u}^{\omega_i,n}_m, \overline{u}^{\omega_i,n}_{m+1}) \qquad \text{and} \qquad \widetilde{u}^{*,+}_{m+\frac{1}{2}} \in [u^{m+1}_{\min}, u^{m+1}_{\max}] \supset \mathscr{I}(\overline{u}^{\omega_i,n}_m, \overline{u}^{\omega_i,n}_{m+1}).$$

Let us start first with $\widetilde{u}^{*,-}_{m+\frac{1}{2}}$.

Case 1. If $\widetilde{u}^{*,-}_{m+\frac{1}{2}} = u^{*}_{m+\frac{1}{2}} - \frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} > u^{m}_{\max}$, we have

$$\theta_{m+\frac{1}{2}} \leq -\frac{\left(u^{m}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, \quad \Delta\mathscr{F}_{m+\frac{1}{2}} < 0.$$

Case 2. If $\widetilde{u}^{*,-}_{m+\frac{1}{2}} = u^{*}_{m+\frac{1}{2}} - \frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} < u^{m}_{\min}$, we have

$$\theta_{m+\frac{1}{2}} \leq \frac{\left(u^{*}_{m+\frac{1}{2}} - u^{m}_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, \quad \Delta\mathscr{F}_{m+\frac{1}{2}} > 0.$$

This way, we have that if $\Delta\mathscr{F}_{m+\frac{1}{2}} \notin \left[-\left(u^{m}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}, \left(u^{*}_{m+\frac{1}{2}} - u^{m}_{\min}\right)\gamma_{m+\frac{1}{2}}\right]$, then

$$\theta_{m+\frac{1}{2}} \leq \begin{cases} -\frac{\left(u^{m}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, & \text{if } \Delta\mathscr{F}_{m+\frac{1}{2}} < 0, \\ \frac{\left(u^{*}_{m+\frac{1}{2}} - u^{m}_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, & \text{if } \Delta\mathscr{F}_{m+\frac{1}{2}} > 0. \end{cases}$$

Let us now consider $\widetilde{u}^{*,+}_{m+\frac{1}{2}}$ and repeating what we did above.

Case 1. If $\widetilde{u}^{*,+}_{m+\frac{1}{2}} = u^{*}_{m+\frac{1}{2}} + \frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} > u^{m+1}_{\max}$, we have

$$\theta_{m+\frac{1}{2}} \leq \frac{\left(u^{m+1}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, \quad \Delta\mathscr{F}_{m+\frac{1}{2}} > 0.$$

Case 2. If $\widetilde{u}^{*,+}_{m+\frac{1}{2}} = u^{*}_{m+\frac{1}{2}} + \frac{\Delta\mathscr{F}_{m+\frac{1}{2}}}{\gamma_{m+\frac{1}{2}}} < u^{m+1}_{\min}$, we have

$$\theta_{m+\frac{1}{2}} \leq -\frac{\left(u^{*}_{m+\frac{1}{2}} - u^{m+1}_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}, \quad \Delta\mathscr{F}_{m+\frac{1}{2}} < 0.$$

Now we need to separate the cases for $\Delta\mathscr{F}_{m+\frac{1}{2}}$.

Case 1. If $\Delta\mathscr{F}_{m+\frac{1}{2}} > 0$,

– If $\Delta\mathscr{F}_{m+\frac{1}{2}} > \left(u^{m+1}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}$, then

$$\theta_{m+\frac{1}{2}} \leq \frac{\left(u^{m+1}_{\max} - u^{*}_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}};$$

– If $\Delta\mathscr{F}_{m+\frac{1}{2}} > \left(u^*_{m+\frac{1}{2}} - u^m_{\min}\right)\gamma_{m+\frac{1}{2}}$, then

$$\theta_{m+\frac{1}{2}} \leq \frac{\left(u^*_{m+\frac{1}{2}} - u^m_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}.$$

This implies that if $\Delta\mathscr{F}_{m+\frac{1}{2}} > 0$, and if $\Delta\mathscr{F}_{m+\frac{1}{2}} > \min\left(u^*_{m+\frac{1}{2}} - u^m_{\min}, u^*_{m+\frac{1}{2}} - u^m_{\min}\right)\gamma_{m+\frac{1}{2}}$, then

$$\theta_{m+\frac{1}{2}} \leq \frac{\min\left(u^*_{m+\frac{1}{2}} - u^m_{\min}, u^*_{m+\frac{1}{2}} - u^m_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}},$$

therefore we will take

$$\theta_{m+\frac{1}{2}} = \min\left(1, \frac{\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}\min\left(u^*_{m+\frac{1}{2}} - u^m_{\min}, u^*_{m+\frac{1}{2}} - u^m_{\min}\right)\right), \quad \Delta\mathscr{F}_{m+\frac{1}{2}} > 0.$$

Case 2. If $\Delta\mathscr{F}_{m+\frac{1}{2}} < 0$,

– If $\Delta\mathscr{F}_{m+\frac{1}{2}} < -\left(u^m_{\max} - u^*_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}$, then

$$\theta_{m+\frac{1}{2}} \leq -\frac{\left(u^m_{\max} - u^*_{m+\frac{1}{2}}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}};$$

– If $\Delta\mathscr{F}_{m+\frac{1}{2}} < -\left(u^*_{m+\frac{1}{2}} - u^{m+1}_{\min}\right)\gamma_{m+\frac{1}{2}}$, then

$$\theta_{m+\frac{1}{2}} \leq -\frac{\left(u^*_{m+\frac{1}{2}} - u^{m+1}_{\min}\right)\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}.$$

Therefore we will take

$$\theta_{m+\frac{1}{2}} = \min\left(1, -\frac{\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}\min\left(u^m_{\max} - u^*_{m+\frac{1}{2}}, u^*_{m+\frac{1}{2}} - u^{m+1}_{\min}\right)\right), \quad \Delta\mathscr{F}_{m+\frac{1}{2}} < 0.$$

Gathering all the conditions gives us the final condition on $\theta_{m+\frac{1}{2}}$ such that our solution conserves max principle, i.e.

$$\theta_{m+\frac{1}{2}} = \begin{cases} \min\left(1, \frac{\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}\min\left(u^*_{m+\frac{1}{2}} - u^m_{\min}, u^{m+1}_{\max} - u^*_{m+\frac{1}{2}}\right)\right), & \text{if } \Delta\mathscr{F}_{m+\frac{1}{2}} > 0, \\ 1, & \text{if } \Delta\mathscr{F}_{m+\frac{1}{2}} = 0, \\ \min\left(1, -\frac{\gamma_{m+\frac{1}{2}}}{\Delta\mathscr{F}_{m+\frac{1}{2}}}\min\left(u^*_{m+\frac{1}{2}} - u^m_{\max}, u^{m+1}_{\min} - u^*_{m+\frac{1}{2}}\right)\right), & \text{if } \Delta\mathscr{F}_{m+\frac{1}{2}} < 0, \end{cases}$$

In fact, $\theta_{m+\frac{1}{2}}$ is useful to gather any convex property we want, and not only conservation of maximum principle but also entropy, water height positivity for Shallow-Water etc...

3.3.3 Numerical validations.

This section is useful to illustrate the robustness and efficiency of our method ; it is not exhaustive though and the reader who want more details can refer to [4] and future François Vilar's work on a priori method.

Linear advection of a composite signal.

To test our stabilization method, let us now address the classical case of the linear advection of a composite signal, composed by the succession of a Gaussian, rectangular, triangular and parabolic signals.

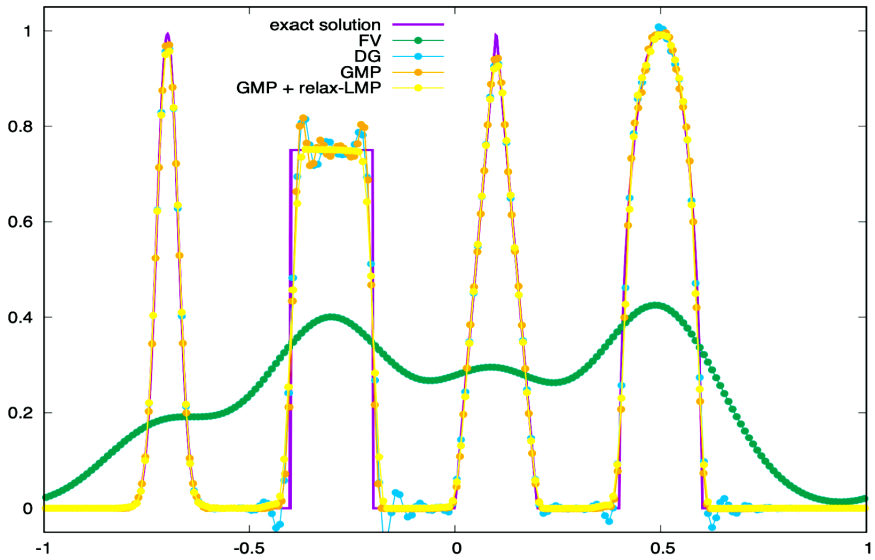


Figure 3.3: $\mathbb{P}^5$-dG/FV sub-mean values on 40 cells for linear advection.

One can see that even by means of this very coarse grid, the numerical solution is extremely precise and robust.

Buckley non-convex flux problem.

We make use of the challenging *Buckley problem* to illustrate some well known problems of discontinuous Galerkin schemes, as entropy and aliasing issues. The Buckley equations writes as

$$\partial_t u + \partial_x F(u) = 0,$$

where the non-convex flux is $F(u) := \frac{4u^2}{4u^2+(1-u)^2}$. Since the flux function is now a complex rational function, it is not practical to analytically integrate the volume integrals in dG schemes. Consequently, we may use as often a quadrature rule, exact for polynomials up to $2k$.

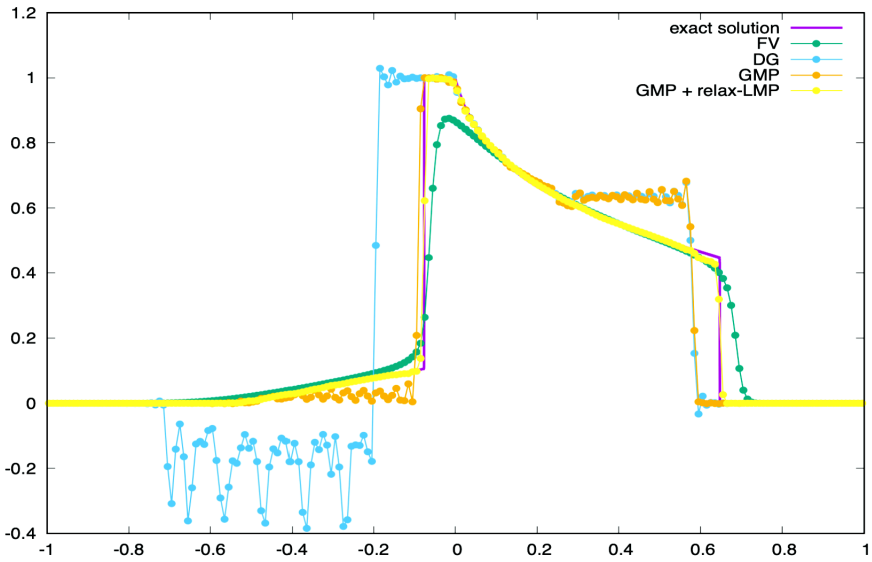


Figure 3.4: $\mathbb{P}^4$ dG/FV sub-mean values on 40 cells for Buckley flux.

We can see here that the classical dG solution does not converge to the entropic one due to aliasing problems : we can then see that our corrected solution prevent this phenomenon.

Section 4

Monolithic subcell convex property preserving scheme for Shallow-Water.

The main goal of this work is to consider the stabilization method introduced previously on Shallow-Water equations. We want to introduce these methods on these equations such that we can pursue Ali Haidar's PhD work. Indeed, one of the objectives will be to consider this correction on NSW coupled with an object floating on the surface. This coupling will be carried out in particular with an ALE method, but this time the two-dimensional case, i.e. the concrete case, will be treated. In this work we will restrain only to the one-dimensional case without any coupling.

4.1 Subcell dG/FV formulation for Shallow-Water.

Let us introduce the L^2-projections of the flux function $\mathbf{F}_h = p^k_{\mathscr{T}_h}(\mathbf{F}(\mathbf{v}_h, b_h))$ and of the source term $\mathbf{B}_h = p^k_{\mathscr{T}_h}(\mathbf{B}(\mathbf{v}_h, \partial_x b_h))$ such that for any $\varphi_{\mathbf{F}}, \varphi_{\mathbf{B}} \in \mathbb{P}^k(\mathscr{T}_h)$, we have

$$\int_{\mathscr{T}_h} \mathbf{F}(\mathbf{v}_h, b_h)\varphi_{\mathbf{F}} = \int_{\mathscr{T}_h} \mathbf{F}_h \varphi_{\mathbf{F}},$$
$$\int_{\mathscr{T}_h} \mathbf{B}(\mathbf{v}_h, \partial_x b_h)\varphi_{\mathbf{B}} = \int_{\mathscr{T}_h} \mathbf{B}_h \varphi_{\mathbf{B}}.$$

As usual, we will consider $\mathscr{F}$ as the interface numerical flux function. Here though, denoting by $\mathbf{v}^-_{i+\frac{1}{2}}$ and $\mathbf{v}^+_{i+\frac{1}{2}}$ respectively the left and right traces of $\mathbf{v}_h$ on interfaces $x_{i+\frac{1}{2}}$ and by $b_{i+\frac{1}{2}} = b^+_{i+\frac{1}{2}} = b^-_{i+\frac{1}{2}}$ the trace of b_h, we define our flux on $x_{i+\frac{1}{2}}$ as follows :

$$\mathscr{F}_{i+\frac{1}{2}} := \mathscr{F}(\mathbf{v}^-_{i+\frac{1}{2}}, \mathbf{v}^+_{i+\frac{1}{2}}, b_{i+\frac{1}{2}}),$$

where we chose the global Lax-Friedrichs flux, i.e.

$$\mathscr{F}(\mathbf{v}^-, \mathbf{v}^+, b) := \frac{\mathbf{F}(\mathbf{v}^-) + \mathbf{F}(\mathbf{v}^-)}{2} - \frac{\sigma}{2}(\mathbf{v}^+ - \mathbf{v}^-),$$

with

$$\sigma := \max_{m\in[\![1,k+1]\!]} \sigma_\omega = \max_{m\in[\![1,k+1]\!]} \left(\max_{m\in[\![1,k+1]\!]} \left[|\overline{u}^\omega_m| + \sqrt{g\overline{H}^\omega_m} \right] \right).$$

From the section (2.2), we can write

$$\int_{\mathscr{T}_h} \partial_t \mathbf{v}_h \varphi - \sum_{\omega\in\mathscr{T}_h} \int_\omega \mathbf{F}_h \partial_x \varphi + \sum_{\omega\in\mathscr{T}_h} \left[\varphi \mathscr{F}\right]_{\partial\omega} = \int_{\mathscr{T}_h} \mathbf{B}_h \varphi, \qquad \forall \varphi \in \mathbb{P}^k(\mathscr{T}_h),$$

or equivalently, using an integration by parts :

$$\int_{\mathscr{T}_h} \partial_t \mathbf{v}_h \varphi + \sum_{\omega\in\mathscr{T}_h} \int_\omega \partial_x \mathbf{F}_h \varphi - \sum_{\omega\in\mathscr{T}_h} \left[\varphi(\mathbf{F}_h - \mathscr{F})\right]_{\partial\omega} = \int_{\mathscr{T}_h} \mathbf{B}_h \varphi, \qquad \forall \varphi \in \mathbb{P}^k(\mathscr{T}_h). \tag{4.1}$$

Substituting $\phi^\omega_m = p^k_\omega(\mathbf{1}^\omega_m)$ defined in (3.1) into (4.1), gives us the local equations on $\omega \in \mathscr{T}_h$:

$$\int_\omega \partial_t \mathbf{v}_\omega \phi^\omega_m = -\int_\omega \partial_x \mathbf{F}_\omega \phi^\omega_m + \int_\omega \mathbf{B}_\omega \phi^\omega_m + \left[\phi^\omega_m(\mathbf{F}_\omega - \mathscr{F})\right]_{\partial\omega}, \qquad \forall m \in [\![1, k+1]\!].$$

Since $\partial_t \mathbf{v}_\omega, \partial_x \mathbf{F}_\omega, \mathbf{B}_\omega \in (\mathbb{P}^k(\omega))^2$ and using the property of ϕ^ω_m, it follows that

$$\partial_t \overline{\mathbf{v}}_\omega = -\frac{1}{|S^\omega_m|} \left([\mathbf{F}_\omega]_{\partial S^\omega_m} - \left[\phi^\omega_m(\mathbf{F}_\omega - \mathscr{F})\right]_{\partial\omega} \right) + \overline{\mathbf{B}}^\omega_m, \qquad \forall m \in [\![1, k+1]\!],$$

where $\overline{\mathbf{v}}_\omega = (\overline{\eta}_\omega, \overline{q}_\omega)^T$ and $\overline{\mathbf{B}}^\omega_m$ are respectively the mean values of $\mathbf{v}_\omega$ and $\mathbf{B}_\omega$ on the subcell S^ω_m. Let us now introduce the $k+2$ subcells interfaces fluxes $\left\{\widetilde{\mathbf{F}}^\omega_{m+\frac{1}{2}}\right\}_{m\in[\![0,k+1]\!]}$ such that

$$\widetilde{\mathbf{F}}^\omega_{m+\frac{1}{2}} - \widetilde{\mathbf{F}}^\omega_{m-\frac{1}{2}} = [\mathbf{F}_\omega]_{\partial S^\omega_m} - \left[\phi^\omega_m(\mathbf{F}_\omega - \mathscr{F})\right]_{\partial\omega}, \qquad \forall m \in [\![1, k+1]\!],$$

such that we have

$$\partial_t \overline{\mathbf{v}}_\omega = -\frac{1}{|S^\omega_m|} \left(\widetilde{\mathbf{F}}^\omega_{m+\frac{1}{2}} - \widetilde{\mathbf{F}}^\omega_{m-\frac{1}{2}} \right) + \overline{\mathbf{B}}^\omega_m, \qquad \forall m \in [\![1, k+1]\!]. \tag{4.2}$$

The formulation (4.2) can then be seen as a FV-like scheme on subcell S^ω_m.

4.2 Subcell low-order corrected FV fluxes.

4.2.1 Benefits and definitions.

Let us define here the *corrected* first-order FV fluxes $\mathscr{F}^{l/r}_{m\pm\frac{1}{2}}$. Such corrected flux are designed in order to :

1. Ensure the desired robustness properties, in particular we aim at preserving the set of admissible states Θ (2.2) ;

2. Obtain a global discrete formulation which is well-balanced i.e. accurately represent the

equilibrium water depth and preserve the balance between gravitational forces and pressure gradients.

Let's consider $\omega_i \in \mathscr{T}_h$ and any subcell $S_m \in \mathscr{T}_{\omega_i}$ and introduce firstly the *sub-mesh reconstructed interface values* for the topography :

$$\overline{b}_{m+\frac{1}{2}} := \max\left(\overline{b}_m, \overline{b}_{m+1}\right) \quad \text{and} \quad \overline{b}_{m-\frac{1}{2}} := \max\left(\overline{b}_{m-1}, \overline{b}_m\right),$$

where $\overline{b} = \frac{1}{|S_m^\omega|}\int_{S_m^\omega} b_h^i \,\mathrm{d}x$, and considering S_m, the *additional subcell's interfaces* topography values, i.e.

$$\begin{aligned}
\overline{b}_m^{m,\pm} &:= \overline{b}_{m\pm\frac{1}{2}} - \max\left(0, \overline{b}_{m\pm\frac{1}{2}} - \overline{\eta}_m\right), \\
\overline{b}_{m+1}^{m,-} &:= \overline{b}_{m+\frac{1}{2}} - \max\left(0, \overline{b}_{m+\frac{1}{2}} - \overline{\eta}_m\right), \\
\overline{b}_{m-1}^{m,+} &:= \overline{b}_{m-\frac{1}{2}} - \max\left(0, \overline{b}_{m-\frac{1}{2}} - \overline{\eta}_m\right).
\end{aligned}$$

We also introduce *subcell's interfaces reconstructions* for the water height as follows :

$$\overline{H}^{\pm} := \max\left(0, \overline{\eta}_m - \overline{b}_{m\pm\frac{1}{2}}\right),$$

and same goes for the surface elevation and discharge, i.e.

$$\overline{\eta}_m^{m,\pm} := \overline{H}_m^{\pm} + \overline{b}_m^{m,\pm} \quad \text{and} \quad \overline{q}_m^{\pm} := \overline{H}_m^{\pm}\frac{\overline{q}_m}{\overline{H}_m},$$

such that we will get a new subcell's interface value :

$$\overline{\mathbf{v}}_m^{m,\pm} := \left(\overline{\eta}_m^{m,\pm}, \overline{q}_m^{\pm}\right)^T.$$

Using these reconstructed values, we introduce some new FV first-order numerical fluxes on subcell's S_m left and right interfaces :

$$\begin{aligned}
\mathscr{F}_{m+\frac{1}{2}}^{l} &:= \mathscr{F}\left(\overline{\mathbf{v}}_m^{m,+}, \overline{\mathbf{v}}_{m+1}^{m,-}, \overline{b}_m^{m,+}\right) + \begin{pmatrix} 0 \\ g\overline{\eta}_m^{m,+}\left(\overline{b}_m^{m,+} - b_{\widetilde{x}_{m+\frac{1}{2}}}\right) \end{pmatrix}, \\
\mathscr{F}_{m-\frac{1}{2}}^{r} &:= \mathscr{F}\left(\overline{\mathbf{v}}_{m-1}^{m,+}, \overline{\mathbf{v}}_m^{m,-}, \overline{b}_m^{m,-}\right) + \begin{pmatrix} 0 \\ g\overline{\eta}_m^{m,-}\left(\overline{b}_m^{m,-} - b_{\widetilde{x}_{m-\frac{1}{2}}}\right) \end{pmatrix},
\end{aligned}$$

where $b_{\widetilde{x}_{m\pm\frac{1}{2}}}$ are the interpolated polynomial values of b_h at interfaces $\widetilde{x}_{m\pm\frac{1}{2}}$.

4.2.2 Preservation of water height positivity for intermediate Riemann states.

Using Forward-Euler time integration with the following CFL condition :

$$\Delta t^n = \frac{\min_{\omega\in\mathcal{T}_h}\left(\frac{h_\omega}{2k+1}, \min_{S_m^\omega}|S_m^\omega|\right)}{\sigma}, \tag{4.3}$$

we can write the formulation (4.2) with reconstructed fluxes as

$$\overline{\mathbf{v}}_m^{n+1} = \overline{\mathbf{v}}_m^n - \frac{\Delta t}{|S_m^\omega|}\left(\mathscr{F}^l_{m+\frac{1}{2}} - \mathscr{F}^r_{m-\frac{1}{2}}\right) + \Delta t\overline{\mathbf{B}}_m,$$

where we dropped the superscript ω. The same way as before, we add then subtract quantities in order to rewrites the previous formulation as a convex combination. Indeed we get :

$$\begin{aligned}
\overline{\mathbf{v}}_m^{n+1} &= \overline{\mathbf{v}}_m^n - \frac{\Delta t}{|S_m^\omega|}\left(\mathscr{F}^l_{m+\frac{1}{2}} - \mathscr{F}^r_{m-\frac{1}{2}}\right) + \Delta t\overline{\mathbf{B}}_m \pm \frac{2\Delta t}{|S_m^\omega|}\sigma \pm \frac{\Delta t}{|S_m^\omega|}\mathbf{F}(\overline{\mathbf{v}}_m^n)\\
&= \left(1 - \frac{2\Delta t}{|S_m^\omega|}\sigma\right)\overline{\mathbf{v}}_m^n + \frac{\Delta t}{|S_m^\omega|}\sigma\left(\overline{\mathbf{v}}_m^n - \frac{\mathscr{F}^l_{m+\frac{1}{2}} - \mathbf{F}(\overline{\mathbf{v}}_m^n)}{\sigma}\right) + \frac{\Delta t}{|S_m^\omega|}\sigma\left(\overline{\mathbf{v}}_m^n - \frac{\mathscr{F}^r_{m-\frac{1}{2}} - \mathbf{F}(\overline{\mathbf{v}}_m^n)}{\sigma}\right) + \Delta t\overline{\mathbf{B}}_m\\
&= \left(1 - \frac{2\Delta t}{|S_m^\omega|}\sigma\right)\overline{\mathbf{v}}_m^n + \frac{\sigma\Delta t}{|S_m^\omega|}\mathbf{v}^{*,l}_{m+\frac{1}{2}} + \frac{\sigma\Delta t}{|S_m^\omega|}\mathbf{v}^{*,r}_{m-\frac{1}{2}} + \Delta t\overline{\mathbf{B}}_m,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{v}^{*,l}_{m+\frac{1}{2}} &:= \overline{\mathbf{v}}_m^n - \frac{\mathscr{F}^l_{m+\frac{1}{2}} - \mathbf{F}(\overline{\mathbf{v}}_m^n)}{\sigma},\\
\mathbf{v}^{*,r}_{m-\frac{1}{2}} &:= \overline{\mathbf{v}}_m^n - \frac{\mathscr{F}^r_{m-\frac{1}{2}} - \mathbf{F}(\overline{\mathbf{v}}_m^n)}{\sigma}.
\end{aligned}$$

Here, since we are interested only in conservation of water height positivity, we just need to ensure that for all $m \in [\![1, k+1]\!]$, $\overline{\eta}_m^n \geq \overline{b}_m^n$. Dropping also the superscript n, we only need to verify is that for any $m \in [\![1, k+1]\!]$, we have

$$\eta^{*,l}_{m+\frac{1}{2}} = \overline{\eta}_m^n - \frac{\mathscr{F}^{(1),l}_{m+\frac{1}{2}} - \overline{q}_m}{\sigma} \geq \overline{b}_m, \quad \text{and} \quad \eta^{*,r}_{m+\frac{1}{2}} = \overline{\eta}_m^n - \frac{\mathscr{F}^{(1),r}_{m+\frac{1}{2}} - \overline{q}_m}{\sigma} \geq \overline{b}_{m+1},$$

where the fluxes are

$$\begin{aligned}
\mathscr{F}^{(1),l}_{m+\frac{1}{2}} &= \frac{\overline{q}_m^+ + \overline{q}_{m+1}^-}{2} - \frac{\sigma}{2}\left(\overline{\eta}_{m+1}^{m,-} - \overline{\eta}_m^{m,+}\right),\\
\mathscr{F}^{(1),r}_{m+\frac{1}{2}} &= \frac{\overline{q}_m^+ + \overline{q}_{m+1}^-}{2} - \frac{\sigma}{2}\left(\overline{\eta}_{m+1}^{m+1,-} - \overline{\eta}_m^{m+1,+}\right).
\end{aligned}$$

Noticing that $\overline{\eta}_{m+1}^{m,-} - \overline{\eta}_m^{m,+} = \overline{H}_{m+1}^- - \overline{H}_m^+ + \overline{b}_{m+1}^{m,-} - \overline{b}_m^{m,+} = \overline{H}_{m+1}^- - \overline{H}_m^+$, and using the notation $\overline{u}_m := \frac{\overline{q}_m}{\overline{H}_m}$, we can write

$$\begin{aligned}
\eta_{m+\frac{1}{2}}^{*,l} &= \overline{\eta}_m - \frac{\mathscr{F}_{m+\frac{1}{2}}^{(1),l} - \overline{q}_m}{\sigma} = \overline{H}_m + \overline{b}_m - \frac{\mathscr{F}_{m+\frac{1}{2}}^{(1),l} - \overline{q}_m}{\sigma} \\
&= \frac{1}{2}\left(2\overline{H}_m - \overline{H}_m^+ + \overline{H}_{m+1}^-\right) - \frac{1}{2\sigma}\left(\overline{q}_{m+1}^- + \overline{q}_m^+ - 2\overline{q}_m\right) + \overline{b}_m \\
&= \frac{1}{2}\left(2\overline{H}_m - \overline{H}_m^+ + \overline{H}_{m+1}^-\right) - \frac{1}{2\sigma}\left(\overline{H}_{m+1}^-\overline{u}_m + \overline{H}_m^+\overline{u}_m - 2\overline{H}_m\overline{u}_m\right) + \overline{b}_m.
\end{aligned}$$

This way, we just need to ensure that

$$\frac{1}{2}\left(2\overline{H}_m - \overline{H}_m^+ + \overline{H}_{m+1}^-\right) - \frac{1}{2\sigma}\left(\overline{H}_{m+1}^-\overline{u}_m + \overline{H}_m^+\overline{u}_m - 2\overline{H}_m\overline{u}_m\right) \geq 0.$$

By construction, for any $p \in [\![1, k+1]\!]$, $0 \leq \overline{H}_p^{\pm} \leq \overline{H}_p$, so we have easily $2\overline{H}_m - \overline{H}_m^+ + \overline{H}_{m+1}^- \geq 2\overline{H}_m$. For the other term, let us rearrange it into

$$\overline{H}_{m+1}^-\overline{u}_m + \overline{H}_m^+\overline{u}_m - 2\overline{H}_m\overline{u}_m = \left(2\overline{H}_m - \overline{H}_m^+\right)\frac{1 + \overline{u}_m\sigma^{-1}}{2} + \overline{H}_{m+1}^-\frac{1 - \overline{u}_{m+1}\sigma^{-1}}{2},$$

and according to CFL (4.3) and construction of water height, we easily have :

$$2\overline{H}_m - \overline{H}_m^+ \geq \overline{H}_m, \qquad \frac{1 + \overline{u}_m\sigma^{-1}}{2} \geq 0, \qquad \overline{H}_{m+1}^- \geq 0, \qquad \frac{1 - \overline{u}_{m+1}\sigma^{-1}}{2} \geq 0.$$

Gathering all the results give us that $\eta_{m+\frac{1}{2}}^{*,l} \geq \overline{b}_m$, what we wanted. A similar proof goes for the case $\eta_{m+\frac{1}{2}}^{*,r} \geq \overline{b}_{m+1}$.

4.3 Evaluation of numerical fluxes.

4.3.1 Expression of blended fluxes.

The values $\left\{\widetilde{\mathbf{F}}_{m+\frac{1}{2}}^{\omega}\right\}_{m\in[\![0,k+1]\!]}$ are here referred as *blended* fluxes, the same way as previous chapter for our *a priori* stabilization method. Considering a mesh element $\omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \in \mathscr{T}_h$, and setting the first and last blended flux to the classical dG numerical flux values at cell boundaries such as

$$\widetilde{\mathbf{F}}_{\frac{1}{2}}^{\omega_i} := \mathscr{F}_{i-\frac{1}{2}}^{\mathrm{dG}} \qquad \text{and} \qquad \widetilde{\mathbf{F}}_{k+\frac{3}{2}}^{\omega_i} := \mathscr{F}_{i+\frac{1}{2}}^{\mathrm{dG}}$$

the m interior fluxes expression are given by

$$\widetilde{\mathbf{F}}_{m+\frac{1}{2}}^{\omega_i} = \mathscr{F}_{m+\frac{1}{2}}^{\mathrm{FV}} + \Theta_{m+\frac{1}{2}}\left(\widehat{\mathbf{F}}_{m+\frac{1}{2}}^{\omega_i} - \mathscr{F}_{m+\frac{1}{2}}^{\mathrm{FV}}\right),$$

where we take $\Theta_{m+\frac{1}{2}} := \mathrm{diag}(\theta_{m+\frac{1}{2}}, \theta_{m+\frac{1}{2}})$, the first-order FV flux $\mathscr{F}^{\mathrm{FV}}$, and $\widehat{\mathbf{F}}$ the following high-order *reconstructed* flux :

$$\widehat{\mathbf{F}}^{\omega_i}_{m+\frac{1}{2}} = \mathbf{F}_{\omega_i}(\widetilde{x}^{\omega_i}_{m+\frac{1}{2}}) - C^{i-\frac{1}{2}}_{m+\frac{1}{2}}\left(\mathbf{F}_{\omega_i}(x_{i-\frac{1}{2}}) - \mathscr{F}_{i-\frac{1}{2}}\right) - C^{i+\frac{1}{2}}_{m+\frac{1}{2}}\left(\mathbf{F}_{\omega_i}(x_{i+\frac{1}{2}}) - \mathscr{F}_{i+\frac{1}{2}}\right),$$

where more details can be found in Chapter (3) or [2, 4].

4.3.2 Blending coefficient assuring water height positivity.

Let's write the formulation (4.2) with blended fluxes as

$$\overline{\mathbf{v}}^{n+1}_m = \overline{\mathbf{v}}^n_m - \frac{\Delta t}{|S^\omega_m|}\left(\widetilde{\mathbf{F}}_{m+\frac{1}{2}} - \widetilde{\mathbf{F}}_{m-\frac{1}{2}}\right) + \Delta t\overline{\mathbf{B}}_m,$$

where we still dropped the superscript ω. As usual, let us add then subtract quantities in order to rewrites this as a convex combination. We then get the system :

$$\overline{\mathbf{v}}^{n+1}_m = \left(1 - \frac{2\Delta t}{|S^\omega_m|}\sigma\right)\overline{\mathbf{v}}^n_m + \frac{\sigma\Delta t}{|S^\omega_m|}\widetilde{\mathbf{v}}^{*,-}_{m+\frac{1}{2}} + \frac{\sigma\Delta t}{|S^\omega_m|}\widetilde{\mathbf{v}}^{*,+}_{m+\frac{1}{2}},$$

and especially the equation on $\overline{\eta}$, i.e.

$$\overline{\eta}^{n+1}_m = \left(1 - \frac{2\Delta t}{|S^\omega_m|}\sigma\right)\overline{\eta}^n_m + \frac{\sigma\Delta t}{|S^\omega_m|}\widetilde{\eta}^{*,-}_{m+\frac{1}{2}} + \frac{\sigma\Delta t}{|S^\omega_m|}\widetilde{\eta}^{*,+}_{m+\frac{1}{2}}.$$

The same way we did in (3), we have two states $\widetilde{\eta}^{*,\pm}_{m+\frac{1}{2}}$, that can be wrote as

$$\widetilde{\eta}^{*,-}_{m+\frac{1}{2}} = \eta^{*,-}_{m+\frac{1}{2}} - \theta_{m+\frac{1}{2}}\left(\frac{\widehat{\mathbf{F}}^{\omega_i,(1)}_{m+\frac{1}{2}} - \mathscr{F}^{l,(1)}_{m+\frac{1}{2}}}{\sigma}\right),$$
$$\widetilde{\eta}^{*,+}_{m+\frac{1}{2}} = \eta^{*,+}_{m+\frac{1}{2}} + \theta_{m+\frac{1}{2}}\left(\frac{\widehat{\mathbf{F}}^{\omega_i,(1)}_{m+\frac{1}{2}} - \mathscr{F}^{r,(1)}_{m+\frac{1}{2}}}{\sigma}\right),$$

and noticing that for all $m \in [\![1, k+1]\!]$, we have $\eta^{*,-}_{m+\frac{1}{2}} \geq \overline{b}_m$ and $\eta^{*,+}_{m+\frac{1}{2}} \geq \overline{b}_{m+1}$, we can find a blending coefficient $\theta_{m+\frac{1}{2}}$ that ensures preservation of water height positivity. The calculation is similar as (3.3.2), and we get here for the first equation on η the following :

$$\theta_{m+\frac{1}{2}} \leq \min\left(\theta^l_{m+\frac{1}{2}}, \theta^r_{m+\frac{1}{2}}\right),$$

where

$$\theta^l_{m+\frac{1}{2}} \leq \begin{cases} \frac{\sigma\left(\eta^{*,-}_{m+\frac{1}{2}} - \overline{b}_m\right)}{\Delta\mathbf{F}^{l,(1)}_{m+\frac{1}{2}}} & \text{if } \Delta\mathbf{F}^{l,(1)}_{m+\frac{1}{2}} \geq 0, \\ 1 & \text{if } \Delta\mathbf{F}^{l,(1)}_{m+\frac{1}{2}} \leq 0, \end{cases} \quad \text{and} \quad \theta^r_{m+\frac{1}{2}} \leq \begin{cases} \frac{\sigma\left(\overline{b}_{m+1} - \eta^{*,+}_{m+\frac{1}{2}}\right)}{\Delta\mathbf{F}^{r,(1)}_{m+\frac{1}{2}}} & \text{if } \Delta\mathbf{F}^{r,(1)}_{m+\frac{1}{2}} \leq 0, \\ 1 & \text{if } \Delta\mathbf{F}^{r,(1)}_{m+\frac{1}{2}} \geq 0, \end{cases}$$

with the notation $\Delta\mathbf{F}^{l/r,(1)}_{m+\frac{1}{2}} := \widehat{\mathbf{F}}^{\omega_i,(1)}_{m+\frac{1}{2}} - \mathscr{F}^{l/r,(1)}_{m+\frac{1}{2}}$.

Section 5

Conclusion and perspectives.

Progress made during this work.

Beginning the internship period, we first successfully developed a `C++` code for solving scalar conservation laws using discontinuous Galerkin methods. This code served as a foundation for our subsequent research, enabling us to explore numerical techniques for efficiently solving conservation laws. Then, we derived Shallow-Water equations from the incompressible Euler equations, establishing a connection between these two fundamental mathematical models. Then by studying the Finite Volume Subcell correction, particularly focusing on Ali Haidar's work on applying these methods to Shallow-Water equations, we gained valuable insights into the *a posteriori* and *a priori* approaches for tackling complex fluid dynamics problems. Especially, for the *a priori* method, which was the main subject of our work, we proved some inequalities on blending coefficient, allowing us to preserve convex properties such as conservation of maximum principle for SCL or preservation of water-height positivity for SW.

Continuation and implementation.

Building upon the knowledge acquired during the internship, our research extended beyond the initial scope. We embarked on implementing the *a priori* approach within two different software frameworks: our homemade `C++` code and Wavebox, an industrial code developed by Fabien Marche. By integrating the a priori approach into these codes, we aim to enhance their capabilities in solving fluid dynamics problems more accurately and efficiently. This implementation phase will allow us to evaluate the performance and effectiveness of the *a priori* approach in practical scenarios.

Ph.D. objectives.

During the forthcoming Ph.D. program, our focus will be on extending our current research to address the challenges posed by the coupling of equations involving a floating object in two di-

mensions. Building upon the foundational work carried out during the internship, which primarily focused on one-dimensional problems, we aim to construct a comprehensive theoretical model that incorporates the dynamics of a floating object within the context of fluid dynamics.
One of the primary objectives of our future research will be the development and implementation of an a priori correction method tailored specifically to this problem. By adapting the existing correction technique, which was successfully applied in the one-dimensional case, we will refine it to accurately capture the intricacies of the fluid-structure interaction phenomena in a two-dimensional setting. This refined correction method will enable us to effectively handle the coupling between the fluid and the floating object.
To facilitate the numerical simulations and enhance the accuracy of our results, we will employ the Arbitrary Lagrangian Eulerian (ALE) description. By utilizing ALE, we can account for the movement and deformation of the computational mesh, allowing for more precise representation of the fluid-structure interaction. The ALE framework will serve as a crucial tool in handling the complex coupling dynamics between the fluid and the floating object.

Applications.

The upcoming research on the coupling of equations with a floating object in two dimensions has significant implications for renewable energy, specifically in the field of wave energy conversion. By utilizing theoretical and numerical findings, we can especially contribute to modeling and optimizing wave energy converters. The untapped potential of wave energy as a clean and abundant resource can be harnessed more effectively, leading to efficient and environmentally friendly wave energy conversion technologies. By providing insights into fluid-structure interactions, we guide the optimization of such devices, improving efficiency and reducing environmental impact.

Section A

Discontinuous Galerkin method for Conservation Laws.

In this appendix we introduce Discontinous Galerkin methods (dG) : the first part of the internship consisted in becoming familiar with these methods and writing a small code in `C++`. Here's some early work on it and a brief introduction to dG implementation.
Our aspiration is to further enhance and refine the code. By expanding its functionality and making it more comprehensive, we aim to create a valuable resource for students seeking a simplified example of dG schemes. This endeavor stems from my own experiences as a student, where access to such a resource would have greatly facilitated my understanding and learning process.
Our code is available on Git, see here : SimpleDG4SCL..

A.1 Hyperbolic scalar conservation law.

We will use dG method on a typical hyperbolic PDE, the scalar conservation law

$$\partial_t u(t,x) + \partial_x f(u(t,x)) = 0, \qquad t \in \mathbb{R}_+, \;\; x \in [a,b]. \tag{A.1}$$

$$u(0,x) = u_0(x), \qquad x \in [a,b]. \tag{A.2}$$

where f is the flux function and u the solution.

A.1.1 Discretization and properties.

Let us introduce some notations and make some assumptions to deal with the problem.

- We consider the domain $[a,b]$ covered by the following mesh composed of the cells $I_i := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, for $i \in [\![1,N]\!]$, where $x_0 := a$ and $x_{N+\frac{1}{2}} := b$.
- We denote $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ for $i \in [\![1,N]\!]$, and $h := \max_{i\in[\![1,N]\!]} \Delta x_i$.

– We assume the mesh is regular, namely there is a constant $c > 0$ independent of h such that $\Delta x_i \geq ch$, for $i \in [\![1, N]\!]$.

– We define a finite element space consisting of piecewise polynomials

$$V_h^k := \{v \in L^2([a,b]) \mid v_{|I_i} \in \mathbb{P}^k(I_i),\ i \in [\![1,N]\!]\},$$

where $\mathbb{P}^k(I_i)$ denotes the set of polynomials of degree up to k defined on the cell I_i.

To get the discrete variational formulation, we multiply each side of the conservation law by a test-function $v_h \in V_h^k$, then integrate on a cell :

$$\partial_t u_h + \partial_x f(u_h) = 0 \Leftrightarrow \int_{I_i} \partial_t u_h v_h \, \mathrm{d}x + \int_{I_i} \partial_x f(u_h) v_h \, \mathrm{d}x = 0 \tag{A.3}$$

$$\Leftrightarrow \int_{I_i} \partial_t u_h v_h \, \mathrm{d}x - \int_{I_i} \partial_x v_h f(u_h) \, \mathrm{d}x + \widehat{f}_{i+\frac{1}{2}} v_h(x^-_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} v_h(x^+_{i-\frac{1}{2}}) = 0, \tag{A.4}$$

where $\widehat{f}$ is the numerical flux which is a single valued function defined at the cell interfaces and in general depends on the values of the numerical solution from both sides of the interface

$$\widehat{f}_{i+\frac{1}{2}} := \widehat{f}\left(u_h(x^-_{i+\frac{1}{2}}, t), u_h(x^+_{i+\frac{1}{2}}, t)\right).$$

From classic finite differences and finite volumes schemes, we have the monotone flux verifying *consistency*, (ie. $\widehat{f}(u,u) = f(u)$), *continuity* ($\widehat{f}$ is at least Lipschitz-continuous with respect to both of its arguments) and *monoticity* ($\widehat{f}$ is a non-decreasing function of its first argument, and a non-decreasing function of its second argument). We can give some examples of well-known monotone fluxes such as *Lax-Friedrichs*'

$$\widehat{f}^{LF}(u^-, u^+) := \frac{f(u^-) + f(u^+)}{2} - \frac{u^- + u^+}{2} \max_u |f'(u)|,$$

or *Godunov* flux defined as

$$\widehat{f}^{God}(u^-, u^+) := \begin{cases} \min_{u^- \leq u \leq u^+} f(u), & \text{if } u^- < u^+ \\ \max_{u^+ \leq u \leq u^-} f(u), & \text{if } u^- \geq u^+ \end{cases}.$$

It is well known that weak solutions to (A.1) are not necessarily unique, and the unique, physically meaningful weak solution (the so-called *entropy solution*) satisfies the following entropy inequality

$$\partial_t U(u) + \partial_x F(u) \leq 0, \tag{A.5}$$

in distribution sense, for any convex entropy $U(u)$ satisfying $U(u) \geq 0$ and the corresponding entropy flux

$$F(u) := \int^u U'(u) f(u) \, \mathrm{d}u.$$

It would be ideal if a numerical approximation to (A.1) has the same entropy inequality as (A.5). A discrete entropy inequality is typically challenging to demonstrate for finite difference or finite volume schemes, particularly for high order schemes and when the flux function f in (A.1) is not convex or concave. However, it turns out that it is easy to prove that the dG scheme (A.4) satisfies a cell entropy inequality.

Proposition A.1.1. *The solution u_h to dG scheme (A.4) satisfies the entropy inequality*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{I_i} U(u_h)\,\mathrm{d}x + \widehat{F}_{i+\frac{1}{2}} - \widehat{F}_{i-\frac{1}{2}} \leq 0, \tag{A.6}$$

for the entropy $U(u) = \frac{u^2}{2}$, and a consistent flux

$$\widehat{F}_{i+\frac{1}{2}} = \widehat{F}\left(u_h(x^-_{i+\frac{1}{2}}, t), u_h(x^+_{i+\frac{1}{2}}, t)\right),$$

such that $\widehat{F}(u,u) = F(u)$.

Proof. We first introduce the short-hand notation

$$B_i(u_h; v_h) := \int_{I_i} \partial_t u_h v_h\,\mathrm{d}x - \int_{I_i} \partial_x v_h f(u_h)\,\mathrm{d}x + \widehat{f}_{i+\frac{1}{2}} v_h(x^-_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} v_h(x^+_{i-\frac{1}{2}}). \tag{A.7}$$

Taking $v_h = u_h$ in B_i, we get

$$\int_{I_i} u_h \partial_t u_h\,\mathrm{d}x - \int_{I_i} \partial_x u_h f(u_h)\,\mathrm{d}x + \widehat{f}_{i+\frac{1}{2}} u_h(x^-_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} u_h(x^+_{i-\frac{1}{2}}) = 0, \tag{A.8}$$

and by denoting $\widetilde{F}(u) = \int^u f(u)\,\mathrm{d}u$, we have

$$\int_{I_i} \partial_t U(u_h)\,\mathrm{d}x - \widetilde{F}\left(u_h(x^-_{i+\frac{1}{2}})\right) + \widetilde{F}\left(u_h(x^+_{i+\frac{1}{2}})\right) + \widehat{f}_{i+\frac{1}{2}} u_h(x^-_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} u_h(x^+_{i-\frac{1}{2}}) = 0, \tag{A.9}$$

or we have

$$B_i(u_h; v_h) = \int_{I_i} \partial_t U(u_h)\,\mathrm{d}x + \widehat{F}_{i+\frac{1}{2}} - \widehat{F}_{i-\frac{1}{2}} + \Theta_{i-\frac{1}{2}},$$

with

$$\widehat{F}_{i+\frac{1}{2}} := -\widetilde{F}\left(u_h(x^-_{i+\frac{1}{2}})\right) + \widehat{f}_{i+\frac{1}{2}} u_h(x^-_{i+\frac{1}{2}}), \tag{A.10}$$

$$\Theta_{i-\frac{1}{2}} := -\widetilde{F}\left(u_h(x^-_{i-\frac{1}{2}})\right) + \widetilde{F}\left(u_h(x^+_{i-\frac{1}{2}})\right) + \widehat{f}_{i+\frac{1}{2}} u_h(x^-_{i-\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} u_h(x^+_{i-\frac{1}{2}}). \tag{A.11}$$

We can verify easily that the numerical entropy flux defined in (A.10) is consistent with the entropy flux $F(u) = \int^u U'(u) f(u)\,\mathrm{d}u$ with $U(u) = \frac{u^2}{2}$. We can also verify that

$$\begin{aligned}\Theta_{i-\frac{1}{2}} &= -\widetilde{F}\left(u_h(x^-_{i-\frac{1}{2}})\right) + \widehat{f}_{i-\frac{1}{2}} u_h(x^-_{i-\frac{1}{2}}) + \widetilde{F}\left(u_h(x^+_{i-\frac{1}{2}})\right) - \widehat{f}_{i-\frac{1}{2}} u_h(x^+_{i-\frac{1}{2}}) \\ &= (u_h(x^+_{i-\frac{1}{2}}) - u_h(x^-_{i-\frac{1}{2}}))(\widetilde{F}'(\xi) - \widehat{f}_{i-\frac{1}{2}}) \geq 0,\end{aligned}$$

where we applied a mean value theorem, ξ being a value between u^- and u^+, and using the fact that $\widetilde{F}'(\xi) = f(\xi)$. We obtain the last inequality with the monoticity of the flux $\widehat{f}$. This finishes the proof. ■

We should remark that the proof is independent of the scheme's accuracy, as it holds for the piecewise polynomial space of any degree k. On any triangulation, an identical proof may be given for a multidimensional dG scheme. Of course, the cell entropy inequality implies an L^2 stability of the numerical solution that we give now.

Proposition A.1.2. *For periodic or compactly supported boundary conditions, the solution u_h to the semi-discrete dG scheme (A.4) satisfies the following*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 u_h^2 \,\mathrm{d}x \leq 0, \quad \text{or} \quad \|u_h(\cdot,t)\|_{L^2} \leq \|u_h(\cdot,0)\|_{L^2}.$$

Proof. We just have to sum up the cell entropy inequality (A.6) over i. The flux terms telescope each other and there is no boundary term left because of the periodic or compact supported boundary condition. The two inequalities are immediate. ■

We can notice that both the cell entropy inequality and the L^2 stability are valid even when the exact solution of the conservation law is discontinuous.

A.1.2 Implementation.

Let us consider the previous spatial discretization with the time discretization with Δt being fixed and $t_n = n\Delta t$. We consider the following semi-discrete variational formulation :

$$\text{Find } u_h \in V_h^k \text{ such that } \int_{I_i} \partial_t u_h \phi \,\mathrm{d}x - \int_{I_i} \partial_x \phi f(u_h) \,\mathrm{d}x + \widehat{f}_{i+\frac{1}{2}}\phi(x_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}}\phi(x_{i-\frac{1}{2}}) = 0, \ \forall \phi \in V_h^k.$$

As we said, in comparison with the classical FEM, here the basis functions (i.e. the basis of V_h^k), are discontinuous across each element. A common choice of basis functions are either the *Lagrange* or the *Legendre* polynomials, because of their orthogonality properties which simplify the computations. In our case we will consider the Legendre polynomials up to degree k as our basis functions. In the reference element $[-1,1]$, the Legendre polynomials are defined recursively as follows

$$\begin{cases} P_0(x) = 1, \ P_1(x) = x, \\ mP_n(x) = (2m-1)xP_{m-1}(x) - (m-1)P_{m-2}(x), \end{cases}$$

for all $m \in \mathbb{N}^* \setminus \{1\}$. To define them on each element, we consider the following map

$$\xi \in [-1,1] \longmapsto \frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2} = x \in I_i. \tag{A.12}$$

So the Legendre polynomials on I_i are denoted $\widetilde{P}_j$, with $j \in [\![0,k]\!]$:

$$\widetilde{P}_j(x) = \widetilde{P}_j\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}\right).$$

The approximate solution u_h can then be wrote as

$$u_{h|I_i} = \sum_{j=0}^{k} u_j(t)\widetilde{P}_j(x),$$

with the $(u_j)_{j\in[\![0,k]\!]}$ that we still want to find. Replacing $u_{h|I_i}$ by its expression in what's above, and considering the test function $\phi = \widetilde{P}_l$, for all $l \in [\![0,k]\!]$, we get

$$\int_{I_i} \frac{\partial}{\partial t}\left(\sum_{j=0}^{k} u_j(t)\widetilde{P}_j(x)\right)\widetilde{P}_l(x)\,\mathrm{d}x = \int_{I_i} f(u_h)\widetilde{P}_l'(x)\,\mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}})$$

$$\Leftrightarrow \sum_{j=0}^{k} \dot{u}_j(t)\left(\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x\right) = \int_{I_i} f(u_h)\widetilde{P}_l'(x)\,\mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}}).$$

We can obtain a fully discrete scheme by using an *Euler-Explicit* approximation of $\dot{u}_j$:

$$\partial_t u(x,t_n)_{|I_i} \simeq \frac{U^{n+1} - U^n}{\Delta t}.$$

Then we simplify the equation

$$\sum_{j=0}^{k} \frac{U_j^{i,n+1} - U_j^{i,n}}{\Delta t}\left(\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x\right) = \int_{I_i} f\left(\sum_{j=0}^{k} U_j^{i,n}\widetilde{P}_j(x)\right)\widetilde{P}_l'(x)\,\mathrm{d}x + \widehat{f}^n_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}^n_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}})$$

$$\Leftrightarrow \sum_{j=0}^{k} U_j^{i,n+1}\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x = \sum_{j=0}^{k} U_j^{i,n}\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x$$
$$+ \Delta t\left(\int_{I_i} f(U^n_{|I_i})\widetilde{P}_l'(x)\,\mathrm{d}x + \widehat{f}^n_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}^n_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}})\right).$$

We finally obtain the variational formulation :

Find $u_h \in V_h^k$ such that, for all $l \in [\![0,k]\!]$,

$$\sum_{j=0}^{k} U_j^{i,n+1}\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x = \sum_{j=0}^{k} U_j^{i,n}\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x$$
$$+ \Delta t\left(\int_{I_i} f(U^n_{|I_i})\widetilde{P}_l'(x)\,\mathrm{d}x + \widehat{f}^n_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}^n_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}})\right),$$

where $\widehat{f}^n_{i+\frac{1}{2}} = \widehat{F}\left(U^n_{|I_{i+1}}, U^n_{|I_i}\right)$ and $U^n_{|I_i} = \sum_{j=0}^k U_j^{i,n}\widetilde{P}_j$ the approximation on I_i at time t_n. We will solve a linear system on each element, as FEM. Therefore for $i \in [\![1, N]\!]$, we can write

$$\sum_{j=0}^k U_j^{i,n+1}\int_{I_i}\widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x = \sum_{j=0}^k U_j^{i,n}\int_{I_i}\widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x + \Delta t\left(\int_{I_i} f(U^n_{|I_i})\widetilde{P}'_l(x)\,\mathrm{d}x + \widehat{f}^n_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}^n_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}})\right)$$

as the scheme

$$\mathsf{G}^i\mathbf{U}^{i,n+1} = \mathsf{G}^i\mathbf{U}^{i,n} + \Delta t\left(\mathbf{S}^{i,n} + \widehat{f}^n_{i-\frac{1}{2}}\mathbf{B}^{i,-} - \widehat{f}^n_{i+\frac{1}{2}}\mathbf{B}^{i,+}\right), \tag{A.13}$$

where G^i is the *mass* matrix, $\mathbf{S}^{i,n}$ the *stiffness* vector, $\mathbf{U}^{i,n}$ the vector of unknowns and $\mathbf{B}^{i,\pm}$ the basis evaluations vector at time step t_n.

$$\mathsf{G}^i = \begin{pmatrix} \int_{I_i}\widetilde{P}_0(x)\widetilde{P}_0(x)\,\mathrm{d}x & \dots & \int_{I_i}\widetilde{P}_0(x)\widetilde{P}_k(x)\,\mathrm{d}x \\ \vdots & \ddots & \vdots \\ \int_{I_i}\widetilde{P}_k(x)\widetilde{P}_0(x)\,\mathrm{d}x & \dots & \int_{I_i}\widetilde{P}_0(x)\widetilde{P}_k(x)\,\mathrm{d}x \end{pmatrix}, \qquad \mathbf{U}^{i,n} = \begin{pmatrix} U_1^{i,n} \\ \vdots \\ U_k^{i,n} \end{pmatrix},$$

$$\mathbf{S}^{i,n} = \begin{pmatrix} \int_{I_i} f(U^n_{|I_i})\widetilde{P}'_0(x)\,\mathrm{d}x \\ \vdots \\ \int_{I_i} f(U^n_{|I_i})\widetilde{P}'_k(x)\,\mathrm{d}x \end{pmatrix}, \qquad \mathbf{B}^{i,\pm} = \left(\widetilde{P}_0(x_{i\pm\frac{1}{2}}), \dots, \widetilde{P}_k(x_{i\pm\frac{1}{2}})\right)^T.$$

In practice we don't compute the matrix form (A.13) on an arbitrary element, we would rather consider the matrix on the reference element $[-1, 1]$ in order to use properties of basis functions, and transform them on each element.

Mass matrix G^i.

Let us consider G^i on I_i. For $i \in [\![1, N]\!]$, we have, for $(j, l) \in [\![1, k]\!]^2$, the following :

$$(\mathsf{G}^i)_{j,l} = \int_{I_i}\widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x = \frac{\Delta x}{2}\int_{-1}^1 \widetilde{P}_j\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}\right)\widetilde{P}_l\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}\right)\,\mathrm{d}\xi.$$

We can remark that (A.12) shifts the domain of $\widetilde{P}_j$ on the right or the left, though the values doesn't change, meaning that we can write

$$\widetilde{P}_j(x) = \widetilde{P}_j\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2}\right) = P_j(\xi).$$

By the orthogonality properties of Legendre polynomials, for all $(j,l) \in [\![1,k]\!]^2$ we have

$$\begin{aligned}(\mathsf{G}^i)_{j,l} &= \frac{\Delta x}{2}\int_{-1}^{1} \widetilde{P}_j\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}\right)\widetilde{P}_l\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}\right)\,\mathrm{d}\xi \\ &= \frac{\Delta x}{2}\int_{-1}^{1} P_j(\xi)P_l(\xi)\,\mathrm{d}\xi = \frac{\Delta x}{2}\frac{2\delta_{jl}}{2j+1},\end{aligned}$$

where the *Kronecker* symbol $\delta_{jl} = \frac{2}{2j+1}$ if $j = l$, and is null otherwise. Thus, the mass matrix becomes diagonal and can be wrote as

$$\mathsf{G}^i = \frac{\Delta x}{2}\,\mathrm{diag}\left(\frac{2}{2j+1}\right)_{j\in[\![1,k]\!]} = \frac{\Delta x}{2}\,\mathrm{diag}\left(2,\frac{2}{3},\ldots,\frac{2}{2k+1}\right).$$

Let us note that for an uniform mesh, the mass matrix is the same for any elements and any time step. Also, we can calculate the coefficients of G^i even if the basis functions aren't orthogonal (using classical quadratures).

Stiffness vector $\mathbf{S}^{i,n}$.

Let us do similarly for $\mathbf{S}^{i,n}$. For $l \in [\![1,k]\!]$, we have

$$\begin{aligned}(\mathbf{S}^{i,n})_l = \int_{I_i} f(U^n_{|I_i})\widetilde{P}_0'(x)\,\mathrm{d}x &= \frac{\Delta x}{2}\int_{-1}^{1} f(U^n_{|I_i})\widetilde{P}_l'\left(\frac{\Delta x}{2}\xi + \frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}\right)\frac{2\,\mathrm{d}\xi}{\Delta x} \\ &= \int_{-1}^{1} f\left(\sum_{j=0}^{k} U_j^{i,n}P_j(\xi)\right)\widetilde{P}_l'(\xi)\,\mathrm{d}\xi \\ &= \int_{-1}^{1} f\left(U^n|_{I_i}^{\xi}\right)\widetilde{P}_l'(\xi)\,\mathrm{d}\xi,\end{aligned}$$

where $U^n|_{I_i}^{\xi} = \sum_{j=0}^{k} U_j^{i,n}P_j(\xi) = U^n_{|I_i}$. Here we will use *Gauss-Legendre* quadrature, obtaining

$$(\mathbf{S}^{i,n})_l = \int_{-1}^{1} f\left(U^n|_{I_i}^{\xi}\right)\widetilde{P}_l'(\xi)\,\mathrm{d}\xi \simeq \sum_{q=1}^{k+1} w_q f\left(U^n|_{I_i}^{\xi_q}\right)P_l'(\xi_q),$$

and obviously the stiffness vector will not be the same depending on the element or the time step, unlike the mass matrix.

Basis evaluations vector $\mathbf{B}^{i,\pm}$.

Finally, it's easy to see that

$$\mathbf{B}^{i,\pm} = \left(\widetilde{P}_0(x_{i\pm\frac{1}{2}}),\ldots,\widetilde{P}_k(x_{i\pm\frac{1}{2}})\right) = \left(P_0(\pm1),\ldots,P_k(\pm1)\right),$$

and using the properties of the basis functions, we get

$$\mathbf{B}^{i,+} = \big(1,\ldots,1\big), \qquad \mathbf{B}^{i,-} = \big(1,-1,\ldots,(-1)^{k-1},(-1)^k\big).$$

Explicit matrix form.

Using again the mapping (A.12), we can write

$$u_{h|I_i} = \sum_{j=0}^{k} U_j^{i,n} \widetilde{P}_j(x) = \sum_{j=0}^{k} U_j^{i,n} \widetilde{P}_j \left(\frac{\Delta x}{2} \xi + \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2} \right) = \sum_{j=0}^{k} U_j^{i,n} P_j(\xi), \quad n \in \mathbb{N},\ x \in I_i,\ \xi \in [-1,1].$$

Finally, we have

$$\mathsf{G}^i \mathbf{U}^{i,n+1} = \mathsf{G}^i \mathbf{U}^{i,n} + \Delta t \left(\mathbf{S}^{i,n} + \widehat{f}^n_{i-\frac{1}{2}} \mathbf{B}^{i,-} - \widehat{f}^n_{i+\frac{1}{2}} \mathbf{B}^{i,+} \right)$$

$$\Leftrightarrow \frac{\Delta x}{2} \operatorname{diag}\left(2, \frac{2}{3}, \dots, \frac{2}{2k+1}\right) \begin{pmatrix} U_1^{i,n} \\ \vdots \\ U_k^{i,n} \end{pmatrix} + \Delta t \left[\begin{pmatrix} \sum_{q=1}^{k+1} w_q f\left(U^n|_{I_i}^{\xi_q}\right) P_0'(\xi_q) \\ \vdots \\ \sum_{q=1}^{k+1} w_q f\left(U^n|_{I_i}^{\xi_q}\right) P_k'(\xi_q) \end{pmatrix} + \widehat{f}^n_{i-\frac{1}{2}} \mathbf{B}^{i,-} - \widehat{f}^n_{i+\frac{1}{2}} \mathbf{B}^{i,+} \right]$$

$$= \frac{\Delta x}{2} \operatorname{diag}\left(2, \frac{2}{3}, \dots, \frac{2}{2k+1}\right) \begin{pmatrix} U_1^{i,n+1} \\ \vdots \\ U_k^{i,n+1} \end{pmatrix}.$$

A.1.3 Numerical validations.

To test our approach, we then considered two very simple test cases with the application code we wrote. First one was $u_0(x) := \sin(2\pi x)$.

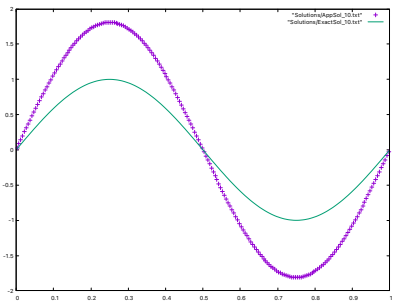


Figure A.1: Order 1 (FV).

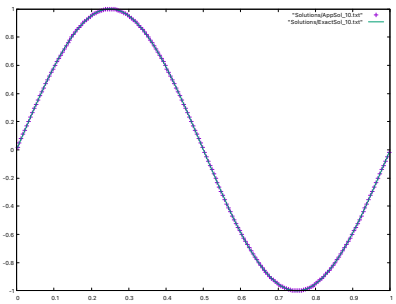


Figure A.2: Order 2.

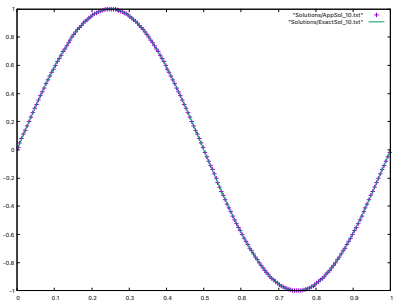


Figure A.3: Order 3.

Second one was the characteristic function $\mathbf{1}_{[0.4,0.6]}$.

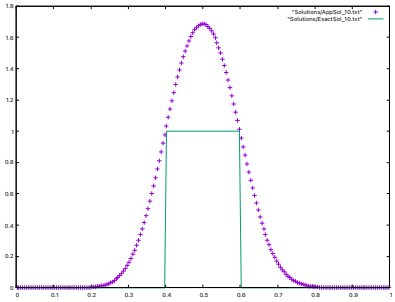


Figure A.4: Order 1 (FV).

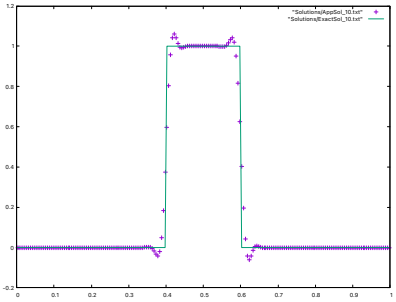


Figure A.5: Order 2.

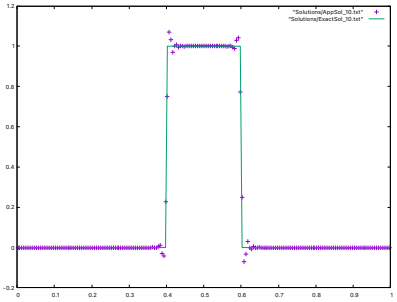


Figure A.6: Order 3.

A.2 Systems of hyperbolic conservation laws.

The discontinuous Galerkin method is easily extendable to the case of systems. The key idea behind this generalisation is to apply the method component-by-component and select numerical fluxes that include all unknown variables. To help with the presentation, we describe the method for a quasilinear system composed of two scalar hyperbolic conservation laws in this section.

Let us consider the following system

$$\begin{cases} \partial_t u(t,x) + \partial_x f(v(t,x)) = 0, & t \in \mathbb{R}_+, \ \ x \in [a,b]. \\ \partial_t v(t,x) + \partial_x g(u(t,x)) = 0, & t \in \mathbb{R}_+, \ \ x \in [a,b]. \end{cases} \tag{A.14}$$

We use the same settings as the previous section, meaning : we keep the domain $[a,b]$ covered by the following mesh composed of the cells $I_i := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, for $i \in [\![1,N]\!]$, where $x_0 := a$ and $x_{N+\frac{1}{2}} := b$. We still denote $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ for $i \in [\![1,N]\!]$, and $h := \max_{i\in[\![1,N]\!]} \Delta x_i$. The mesh is still assumed to be regular. We also still consider the following space

$$V_h^k := \{v \in L^2(\mathbb{R}) \mid v_{|I_i} \in \mathbb{P}^k(I_i), \ i \in [\![1,N]\!]\},$$

where $u_h \in V_h^k$ (resp. $v_h \in V_h^k$) will be the approximation of u (resp. v).

Then, the same way as earlier, we obtain the semi-discrete scheme for both equations, considering a test function $\phi \in V_h^k$ and using (A.14). We finally get

$$\text{Find } (u_h, v_h) \in (V_h^k)^2 \text{ such that, for all } \phi \in V_h^k,$$
$$\begin{cases} \displaystyle\int_{I_i} \partial_t u_h \phi \, \mathrm{d}x - \int_{I_i} \partial_x \phi f(v_h) \, \mathrm{d}x + \widehat{f}_{i+\frac{1}{2}} \phi(x_{i+\frac{1}{2}}) - \widehat{f}_{i-\frac{1}{2}} \phi(x_{i-\frac{1}{2}}) = 0, \\ \displaystyle\int_{I_i} \partial_t v_h \phi \, \mathrm{d}x - \int_{I_i} \partial_x \phi g(u_h) \, \mathrm{d}x + \widehat{g}_{i+\frac{1}{2}} \phi(x_{i+\frac{1}{2}}) - \widehat{g}_{i-\frac{1}{2}} \phi(x_{i-\frac{1}{2}}) = 0, \end{cases}$$

where $\widehat{f}_{i+\frac{1}{2}} = \mathscr{F}\left(u^+_{i+\frac{1}{2}}, u^-_{i+\frac{1}{2}}, v^+_{i+\frac{1}{2}}, v^-_{i+\frac{1}{2}}\right)$ and $\widehat{g}_{i+\frac{1}{2}} = \mathscr{G}\left(u^+_{i+\frac{1}{2}}, u^-_{i+\frac{1}{2}}, v^+_{i+\frac{1}{2}}, v^-_{i+\frac{1}{2}}\right)$ are the numerical fluxes. To obtain the scheme's conservation property, the numerical fluxes $\mathscr{F}$ and $\mathscr{G}$ must meet the qualities referred to on the numerical flux in the scalar case, namely *consistency*, *Lipschitz continuity*, and *monotonicity*.

Remark A.2.1. A numerical flux used for the scalar case can be generalised to the case of systems. In order to obtain the generalised form of the flux, it is reasonable to rewrite the system (A.14) in a compact form, i.e.

$$\mathbf{U}_t + \mathsf{A}(\mathbf{U})\mathbf{U}_x = 0,$$

where $\mathbf{U} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^2 = (u, v)^T$ and $\mathsf{A} \in \mathscr{M}_{2\times 2}(\mathbb{R})$ is a matrix depending on $\mathbf{U}$:

$$\mathsf{A} = \begin{pmatrix} 0 & f'(v) \\ g'(u) & 0 \end{pmatrix}.$$

Let us now suppose that the numerical flux $\widehat{f}_{i+\frac{1}{2}}$ in scalar case is $\widehat{f}_{i+\frac{1}{2}} = \widehat{f}(u^-, u^+)$, where $u^+ = u_h(x^+_{i+\frac{1}{2}}, t)$ and $u^- = u_h(x^-_{i+\frac{1}{2}}, t)$ are the values of the approximate solution on the left and right element respectively where $x^+_{i+\frac{1}{2}}$ is located. Then, the generalised form of the numerical flux is

$$\widehat{f}_{i+\frac{1}{2}} = F(u^+, u^-, v^+, v^-).$$

Following the procedure of the scalar case, we define similarly the basis functions, meaning we consider the Legendre polynomials up to degree k as the basis functions. We can then write the numerical solutions as

$$u_{h|I_i} = \sum_{j=0}^{k} u_j(t)\widetilde{P}_j(x), \qquad v_{h|I_i} = \sum_{m=0}^{k} u_m(t)\widetilde{P}_m(x).$$

Hence, for $(j, m, l) \in [\![0, k]\!]^3$, we obtain

$$\begin{cases} \displaystyle\int_{I_i} \frac{\partial}{\partial t}\left(\sum_{j=0}^{k} u_j(t)\widetilde{P}_j(x)\right)\widetilde{P}_l(x)\,\mathrm{d}x = \int_{I_i} f(v_h)\widetilde{P}'_l(x)\,\mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}}), \\ \displaystyle\int_{I_i} \frac{\partial}{\partial t}\left(\sum_{m=0}^{k} u_m(t)\widetilde{P}_m(x)\right)\widetilde{P}_l(x)\,\mathrm{d}x = \int_{I_i} g(u_h)\widetilde{P}'_l(x)\,\mathrm{d}x + \widehat{g}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{g}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}}). \end{cases}$$

$$\Leftrightarrow \begin{cases} \displaystyle\sum_{j=0}^{k} \dot{u}_j(t)\left(\int_{I_i} \widetilde{P}_j(x)\widetilde{P}_l(x)\,\mathrm{d}x\right) = \int_{I_i} f(v_h)\widetilde{P}'_l(x)\,\mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}}), \\ \displaystyle\sum_{j=0}^{k} \dot{u}_j(t)\left(\int_{I_i} \widetilde{P}_m(x)\widetilde{P}_l(x)\,\mathrm{d}x\right) = \int_{I_i} g(u_h)\widetilde{P}'_l(x)\,\mathrm{d}x + \widehat{g}_{i-\frac{1}{2}}\widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{g}_{i+\frac{1}{2}}\widetilde{P}_l(x_{i+\frac{1}{2}}). \end{cases}$$

For an approximation of the time derivative, we use again the *Forward Euler* method. It follows

that we have

$$\begin{cases} \displaystyle\sum_{j=0}^{k} \frac{U_j^{i,n+1} - U_j^{i,n}}{\Delta t} \left(\int_{I_i} \widetilde{P}_j \widetilde{P}_l \, \mathrm{d}x \right) = \int_{I_i} g \left(\sum_{m=0}^{k} V_m^{i,n} \widetilde{P}_m \right) \widetilde{P}_l' \, \mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}^n \widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}^n \widetilde{P}_l(x_{i+\frac{1}{2}}), \\ \displaystyle\sum_{j=0}^{k} \frac{V_m^{i,n+1} - V_m^{i,n}}{\Delta t} \left(\int_{I_i} \widetilde{P}_m \widetilde{P}_l \, \mathrm{d}x \right) = \int_{I_i} f \left(\sum_{j=0}^{k} U_j^{i,n} \widetilde{P}_j \right) \widetilde{P}_l' \, \mathrm{d}x + \widehat{g}_{i-\frac{1}{2}}^n \widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{g}_{i+\frac{1}{2}}^n \widetilde{P}_l(x_{i+\frac{1}{2}}). \end{cases}$$

$$\Leftrightarrow \begin{cases} \displaystyle\sum_{j=0}^{k} U_j^{i,n+1} \int_{I_i} \widetilde{P}_j \widetilde{P}_l \, \mathrm{d}x = \sum_{j=0}^{k} U_j^{i,n} \int_{I_i} \widetilde{P}_j \widetilde{P}_l \, \mathrm{d}x + \Delta t \left(\int_{I_i} f(V_{|I_i}^n) \widetilde{P}_l' \, \mathrm{d}x + \widehat{f}_{i-\frac{1}{2}}^n \widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{f}_{i+\frac{1}{2}}^n \widetilde{P}_l(x_{i+\frac{1}{2}}) \right), \\ \displaystyle\sum_{j=0}^{k} V_m^{i,n+1} \int_{I_i} \widetilde{P}_m \widetilde{P}_l \, \mathrm{d}x = \sum_{j=0}^{k} V_m^{i,n} \int_{I_i} \widetilde{P}_m \widetilde{P}_l \, \mathrm{d}x + \Delta t \left(\int_{I_i} g(U_{|I_i}^n) \widetilde{P}_l' \, \mathrm{d}x + \widehat{g}_{i-\frac{1}{2}}^n \widetilde{P}_l(x_{i-\frac{1}{2}}) - \widehat{g}_{i+\frac{1}{2}}^n \widetilde{P}_l(x_{i+\frac{1}{2}}) \right), \end{cases}$$

where $U_{|I_i}^n = \sum_{j=0}^k U_j^{i,n} \widetilde{P}_j$ and $V_{|I_i}^n = \sum_{m=0}^k U_m^{i,n} \widetilde{P}_m$ are the approximations of u and v respectively, in I_i at time t_n. Thus, the final linear system can be written in a matrix form such as

$$\begin{cases} \mathsf{G}^i \mathbf{U}_j^{i,n+1} = \mathsf{G}^i \mathbf{U}_j^{i,n} + \Delta t \left((\mathbf{S}_v)^{i,n} + \widehat{f}_{i-\frac{1}{2}}^n \mathbf{B}^{i,-} - \widehat{f}_{i+\frac{1}{2}}^n \mathbf{B}^{i,+} \right), \\ \mathsf{G}^i \mathbf{V}_m^{i,n+1} = \mathsf{G}^i \mathbf{V}_m^{i,n} + \Delta t \left((\mathbf{S}_u)^{i,n} + \widehat{g}_{i-\frac{1}{2}}^n \mathbf{B}^{i,-} - \widehat{g}_{i+\frac{1}{2}}^n \mathbf{B}^{i,+} \right), \end{cases} \tag{A.15}$$

for every element I_i, $i \in [\![1, N]\!]$. Following the simplifications and the remarks made on previous section, the system (A.15) can easily be computed.

Section B

Derivation of Shallow-Water Equations.

Shallow-Water Equations (SWEs) are a set of mathematical equations used to describe water movement in a shallow region, which is typically defined as a region where the water depth is much smaller than the horizontal scale of the flow. These equations are derived from the more general fluid mechanics equations known as the *Euler Equations*. Let us consider the flow variable

$$\mathbf{V} := \mathbf{V}(x,y,z,t) = \begin{pmatrix} U(x,y,z,t) \\ V(x,y,z,t) \\ W(x,y,z,t) \end{pmatrix}.$$

In this context, let us recall that the incompressibility condition is written as

$$\mathrm{div}(\mathbf{V}) = \partial_x U(x,y,z,t) + \partial_y V(x,y,z,t) + \partial_z W(x,y,z,t) = 0,$$

and the speed evolution is given as

$$\partial_t \mathbf{V} + \mathrm{div}(\mathbf{V} \otimes \mathbf{V}) = \mathbf{g} - \rho^{-1}\boldsymbol{\nabla} p,$$

where $\mathbf{g} = (0,0,-g)^T$, with g the gravity condition. We can re-write the previous system as

$$\begin{cases} \partial_t U + \partial_x U^2 + \partial_y (UV) + \partial_z (UW) = -\rho^{-1}\partial_x p, \\ \partial_t V + \partial_x (UV) + \partial_y V^2 + \partial_z (VW) = -\rho^{-1}\partial_y p, \\ \partial_t W + \partial_x (UW) + \partial_y (V x W) + \partial_z W^2 = -\left(g + \rho^{-1}\partial_z p\right). \end{cases}$$

In practice, the numerical resolution of these equations is prohibitively expensive. On the one hand, it is a system with three unknowns, functions of the three spatial and temporal dimensions. Furthermore, because the system only describes the evolution of the velocity field in the fluid, we lack direct information on the domain of calculation, which can change over time according to the deformations of the free surface, and thus adds to the problem's unknowns[1].

[1]Many works have been done over several decades to return to equations that are easier to study and process

A traditional approach entails integrating these equations on the profunder in order to reduce the dimension of the problem by removing the vertical variable. This technique must be used in conjunction with appropriate boundary conditions and certain assumptions that simplify the flow regimes under consideration. That's why we use SWEs to describe flows of relatively shallow depth, as they provide a simplified representation of fluid motion in Shallow-Water environments, making them a useful tool for a wide range of applications, including oceanography, hydraulic engineering, and meteorology. They are particularly useful for simulating phenomena such as tidal waves, storm surges, and tsunamis, which play a critical role in coastal management and disaster response.

B.1 Boundary conditions.

We first need to precise the boundary conditions who will intervene in the derivation of those equations. For the rest of this section, Let us consider $Z(x, y, t)$ the total elevation of the water surface and $h(x, y, t)$ the water height. We will deal a topography term $b : (x, y) \in \mathbb{R}^2 \mapsto b(x, y)$, supposed to be $\mathscr{C}^1$ so we can have $Z = h + b$.

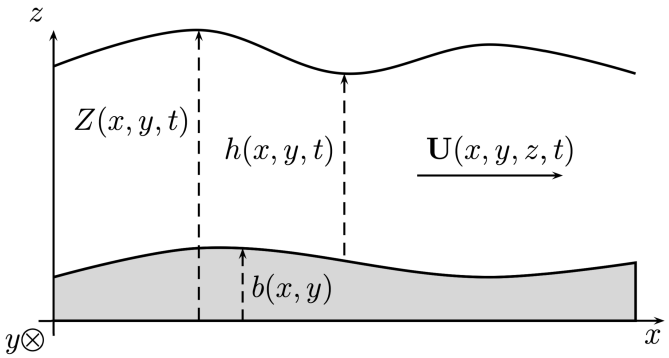


Figure B.1: Representation of flow parameters (from [1]).

We will also note $\mathbf{U} := \mathbf{U}(x, y, z, t) = \begin{pmatrix} U(x, y, z, t) \\ V(x, y, z, t). \end{pmatrix}$ the horizontal speed vector.

B.1.1 Non-penetration condition.

This is to express that the fluid cannot cross the bottom surface. As a result, the speed of a particle at the bottom is null in the direction perpendicular to the surface. Let us consider M_b a point of coordinates $(x_0, y_0, b(x_0, y_0))$ belonging to this surface. For any particule's trajectory on the surface

digitally.

of the bottom, given by the map

$$\boldsymbol{\phi} : \zeta \longmapsto (x(\zeta), y(\zeta), b(x(\zeta), y(\zeta))^T ,$$

such that $\boldsymbol{\phi}(0) = M_b$, we can verify that $\mathbf{n} = \big(-\partial_x b(x_0, y_0), -\partial_y b(x_0, y_0), 1\big)^T$ satisfy $\mathbf{n} \cdot \boldsymbol{\phi}'(0) = 0$, meaning that $\mathbf{n}$ is a normal vector at M_b. The condition $\mathbf{V} \cdot \mathbf{n} = 0$ at point M_b is then

$$W_b = \mathbf{U}_b \cdot \boldsymbol{\nabla} b, \qquad \text{with} \;\; W_b := W(x_0, y_0, b(x_0, y_0), t)) \;\; \text{and} \;\; \mathbf{U}_b := \mathbf{U}(x_0, y_0, b(x_0, y_0), t)).$$

B.1.2 Kinematic condition on the free surface.

This condition expresses the fact that a particle on the free surface does not move freely. Consider the path of a particle

$$\boldsymbol{\phi} : t \longmapsto (x(t), y(t), z(t))^T = (x(t), y(t), Z(x(t), y(t), t))^T .$$

We have the horizontal speed of the particle being $\mathbf{U}_Z = (\dot{x}(t), \dot{y}(t))^T$. We can then deduce from parametrization of $z(t)$ the expression of vertical component of speed $W_z = \dot{z}(t)$ on the surface

$$W_z = \partial_t Z + \mathbf{U}_Z \cdot \boldsymbol{\nabla} Z.$$

B.2 Nondimensionalization of equations.

We are interested in regimes of shallow relative depth, regimes in which the wavelength λ is large in comparison to the characteristic depth h_0, or regimes in which the parameter $\mu = h_0 \lambda^{-1}$ is very small. As a result, the Shallow-Water model, which is based on this approximation, provides a good representation of the mechanisms near the coast[2]. These equations, however, can also be used to

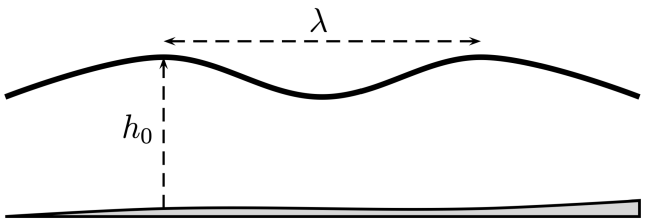


Figure B.2: Wavelengths and characteristic depth (from [1]).

study large-scale oceanographic phenomena, such as tsunamis. In practice, the latter propagate over a few kilometres of depth, whereas the wavelengths involved can reach several hundred kilometers. These propagations are characterized by a small variation of the flow quantities according to the vertical coordinate, which will result in some scaling assumptions in problem scaling, as we will see.

[2] λ is in the order of ten meters long and h_0 is in the order of a meter.

Let us introduce the following quantities :

$$\mu = \frac{h_0}{L} \ ; \ \widetilde{x} = \frac{x}{L} \ ; \ \widetilde{y} = \frac{y}{L} \ ; \ \widetilde{z} = \frac{z}{h_0} \ ; \ \widetilde{t} = \mu t \sqrt{\frac{g}{h_0}} \ ;$$
$$\widetilde{U} = \frac{U}{\sqrt{gh_0}} \ ; \ \widetilde{V} = \frac{V}{\sqrt{gh_0}} \ ; \ \widetilde{W} = \frac{W}{\mu\sqrt{gh_0}} \ ; \ \widetilde{p} = \frac{p}{\rho g h_0}.$$

It is worth noting that the previously mentioned scale factors on the variable vertical z (nondimensioned by $h0 = \mu L$) and the vertical component of the velocity w (nondimensioned by $\mu\sqrt{gh_0}$) can be found here. Let us recall that the parameter is assumed to be small : in the following section, we will neglect the terms of order μ^2. We can verify easily that the incompressibility condition on the adimensioned variables is written

$$\mathrm{div}(\widetilde{\mathbf{V}}) = 0,$$

and that the previous system give us

$$\begin{cases} \partial_{\widetilde{t}}\widetilde{U} + \partial_{\widetilde{x}}\widetilde{U}^2 + \partial_{\widetilde{y}}\widetilde{U}\widetilde{V} + \partial_{\widetilde{z}}\widetilde{U}\widetilde{W} = -\partial_{\widetilde{x}}\widetilde{p}, \\ \partial_{\widetilde{t}}\widetilde{U} + \partial_{\widetilde{x}}\widetilde{U}\widetilde{V} + \partial_{\widetilde{y}}\widetilde{V}^2 + \partial_{\widetilde{z}}\widetilde{V}\widetilde{W} = -\partial_{\widetilde{y}}\widetilde{p}, \\ \mu^2\left(\partial_{\widetilde{t}}\widetilde{W} + \partial_{\widetilde{x}}\widetilde{U}\widetilde{W} + \partial_{\widetilde{y}}\widetilde{V}\widetilde{W} + \partial_{\widetilde{z}}\widetilde{W}^2\right) = -(1 + \partial_{\widetilde{z}}\widetilde{p}). \end{cases}$$

In the following we will ommit the « tilde » symbol because the nondimensionalized boundary conditions are the same as the original ones.

B.3 Mass equation.

First Let us recall a fundamental proposition.

Proposition B.3.1 (Leibniz formula). *For any function $f \in \mathscr{C}^0$, and for any u,v differentiable, we have the following*[3] *:*

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_{u(x)}^{v(x)} f(x,t)\ \mathrm{d}t = \int_{u(x)}^{v(x)} \partial_x f(x,t)\ \mathrm{d}t + v'(x)f(v(x)) - u'(x)f(u(x)).$$

Let us now integrate the incompressibility condition on the depth $[b, Z]$. We get

$$\int_b^Z \mathrm{div}(\mathbf{V})\ \mathrm{d}z = \int_b^Z \mathrm{div}(\mathbf{U})\ \mathrm{d}z + \int_b^Z \partial_z W\ \mathrm{d}z = \mathrm{div}\left(\int_b^Z \mathbf{U}\ \mathrm{d}z\right) - \mathbf{U}_Z \cdot \boldsymbol{\nabla} Z + \mathbf{U}_b \cdot \boldsymbol{\nabla} b + W_Z - W_b,$$

[3]Note that this formula can be generalized for superior dimensions.

with surface speed $(\mathbf{U}_Z, W_Z)$ and bottom speed $(\mathbf{U}_b, W_b)$. For the rest of the section, we will use the shortcut

$$\mathbf{u} = \begin{pmatrix} u \\ \\ v \end{pmatrix} = \fint_b^Z \mathbf{U}\, \mathrm{d}z = \frac{1}{h}\int_b^Z \mathbf{U}\, \mathrm{d}z.$$

for the mean value of $\mathbf{U}$ on the vertical axis.

By using the conditions $W_z = \partial_t Z + \mathbf{U}_Z \cdot \boldsymbol{\nabla} Z$ and $W_b = \mathbf{U}_b \cdot \boldsymbol{\nabla} b$, and remarking that since b is independant of time, such that $\partial_t Z = \partial_t(h+b) = \partial_t h$, we get the following called *mass equation*

$$\partial_t h + \mathrm{div}(h\mathbf{u}) = 0.$$

B.4 Momentum equation.

Let us now integrate over the depth the system obtained earlier. We get

$$\int_b^Z \partial_t U\, \mathrm{d}z + \int_b^Z \partial_x U^2\, \mathrm{d}z + \int_b^Z \partial_y(UV)\, \mathrm{d}z + \int_b^Z \partial_z(UW)\, \mathrm{d}z + \int_b^Z \partial_x p\, \mathrm{d}z = 0.$$

Then we consider each term and obtain

$$\begin{aligned}
\int_b^Z \partial_t U\, \mathrm{d}z &= \frac{\partial}{\partial t}\left(\int_b^Z U\, \mathrm{d}z\right) - U_Z \partial_t Z = \partial_t(hu) - U_Z \partial_t Z, \\
\int_b^Z \partial_x U^2\, \mathrm{d}z &= \frac{\partial}{\partial x}\left(\int_b^Z U^2\, \mathrm{d}z\right) - U_Z^2 \partial_x Z + U_b^2 \partial_x b, \\
\int_b^Z \partial_y(UV)\, \mathrm{d}z &= \frac{\partial}{\partial y}\left(\int_b^Z UV\, \mathrm{d}z\right) - U_Z V_Z \partial_y Z + U_b V_b \partial_y b, \\
\int_b^Z \partial_z(UW)\, \mathrm{d}z &= W_Z U_Z - W_b U_b = U_Z \partial_t Z + U_Z(\mathbf{U}_Z \cdot \boldsymbol{\nabla} Z) - U_b(\mathbf{U}_b \cdot \boldsymbol{\nabla} b).
\end{aligned}$$

Considering those calculations, we get

$$\begin{aligned}
&\int_b^Z \partial_t U \,\mathrm{d}z + \int_b^Z \partial_x U^2 \,\mathrm{d}z + \int_b^Z \partial_y(UV) \,\mathrm{d}z + \int_b^Z \partial_z(UW) \,\mathrm{d}z + \int_b^Z \partial_x p \,\mathrm{d}z \\
&= \partial_t(hu) - U_Z \partial_t Z + \frac{\partial}{\partial x}\left(\int_b^Z U^2 \,\mathrm{d}z\right) - U_Z^2 \partial_x Z + U_b^2 \partial_x b + \frac{\partial}{\partial y}\left(\int_b^Z U^2 \,\mathrm{d}z\right) \\
&- U_Z V_Z \partial_y Z + U_b V_b \partial_y b + U_Z \partial_t Z + U_Z(\mathbf{U}_Z \cdot \boldsymbol{\nabla} Z) - U_b(\mathbf{U}_b \cdot \boldsymbol{\nabla} b) + \int_b^Z \partial_x p \,\mathrm{d}z \\
&= \partial_t(hu) - U_Z \partial_t Z + \frac{\partial}{\partial x}\left(\int_b^Z U^2 \,\mathrm{d}z\right) + \frac{\partial}{\partial y}\left(\int_b^Z UV \,\mathrm{d}z\right) - U_Z \partial_t Z \\
&- U_Z(\mathbf{U}_Z \cdot \boldsymbol{\nabla} Z) + U_b(\mathbf{U}_b \cdot \boldsymbol{\nabla} b) + U_Z \partial_t Z + U_Z(\mathbf{U}_Z \cdot \boldsymbol{\nabla} Z) - U_b(\mathbf{U}_b \cdot \boldsymbol{\nabla} b) + \int_b^Z \partial_x p \,\mathrm{d}z \\
&= \partial_t(hu) + \frac{\partial}{\partial x}\left(\int_b^Z U^2 \,\mathrm{d}z\right) + \frac{\partial}{\partial y}\left(\int_b^Z UV \,\mathrm{d}z\right) + \int_b^Z \partial_x p \,\mathrm{d}z.
\end{aligned}$$

Therefore the equation can be written as

$$\partial_t(hu) + \frac{\partial}{\partial x}\left(\int_b^Z U^2 \,\mathrm{d}z\right) + \frac{\partial}{\partial y}\left(\int_b^Z UV \,\mathrm{d}z\right) + \int_b^Z \partial_x p \,\mathrm{d}z = 0.$$

It then remains to deal with the advective terms and the pressure terms.

B.4.1 Advective terms.

In the literature, we sometimes simply consider a constant speed along the vertical[4], so that $\mathbf{u} = \mathbf{U}$. We'll be more general here, assuming that the flow is *weakly sheared*. In other words, we allow a speed deviation along the vertical axis that is weak enough to be ignored in the considered asymptotic regime :

$$\mathbf{U}(x, y, z, t) = \mathbf{u}(x, y, t) + \mu \mathbf{u}'(x, y, z, t).$$

Using the definition of $\mathbf{u}$, we get $\int_b^Z \mathbf{u}' \,\mathrm{d}z = 0$, and therefore

$$\begin{aligned}
&\int_b^Z U^2 \,\mathrm{d}z = \int_b^Z (u + \mu u')^2 \,\mathrm{d}z = \int_b^Z u^2 \,\mathrm{d}z + \mu^2 \int_b^Z (u')^2 \,\mathrm{d}z + 2\mu u \int_b^Z u' \,\mathrm{d}z = hu^2 + \mathcal{O}(\mu^2), \\
&\int_b^Z UV \,\mathrm{d}z = \int_b^Z (u + \mu u')(v + \mu v') \,\mathrm{d}z = huv + \mathcal{O}(\mu^2).
\end{aligned}$$

We can of course neglect residual terms contained in $\mathcal{O}(\mu^2)$, as we've stated in the introduction of B.2.

[4]This hypothesis may appear harsh, but in reality, it expresses the fact that it is possible to ignore speed variations on the vertical axis.

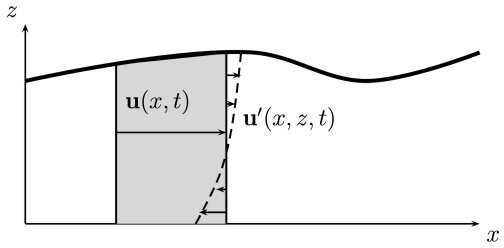


Figure B.3: Decomposition of $\mathbf{U}$ in mean speed $\mathbf{u}$ and horizontal deviation speed $\mathbf{u}'$ (from [1]).

B.4.2 Pressure terms and final formulation.

We recall the last equation from the previous system :

$$\mu^2 \left(\partial_t W + \partial_x UW + \partial_y VW + \partial_z W^2\right) = -(1 + \partial_z p).$$

It gives us, by integrating, the following equation[5]

$$\partial_z p = -1 + \mathcal{O}(\mu^2) \Rightarrow p(Z) - p(z) = z - Z + \mathcal{O}(\mu^2)$$

Assuming the atmospheric pressure is null at the surface (i.e. $p(Z) = 0$), we have immediately that $p(b) = h + \mathcal{O}(\mu^2)$, and we get

$$\int_b^Z p(z)\,\mathrm{d}z = \frac{h^2}{2} + \mathcal{O}(\mu^2).$$

Then we can write

$$\int_b^Z \partial_x p(z)\,\mathrm{d}z = \frac{\partial}{\partial x}\left(\int_b^Z p(z)\,\mathrm{d}z\right) - p(Z)\partial_x Z + p(b)\partial_x b = \frac{\partial}{\partial x}\left(\frac{h^2}{2}\right) + h\partial_x b + \mathcal{O}(\mu^2).$$

Considering the previous work on advective terms and what's above, we get the two *momentum equations* here :

$$\partial_t(hu) + \frac{\partial}{\partial x}\left(hu^2 + \frac{h^2}{2}\right) + \partial_y(huv) = -h\partial_x b,$$
$$\partial_t(hv) + \partial_x(huv) + \frac{\partial}{\partial y}\left(huv + \frac{h^2}{2}\right) = -h\partial_y b.$$

[5] In fact, by coming back to dimensioned variables, still neglecting terms from $\mathcal{O}(\mu^2)$, this equation becomes

$$\partial_z p = -\rho g.$$

This is called the *hydrostatic pressure* hypothesis,assuming a linear pressure distribution along the vertical. It reflects the fact that the pressure at a point (x, y) depends only on the weight of the water column above this point.

The final system is then

$$\begin{aligned}
&\partial_t h + \operatorname{div}(h\mathbf{u}) = 0,\\
&\partial_t(hu) + \frac{\partial}{\partial x}\left(hu^2 + \frac{h^2}{2}\right) + \partial_y(huv) = -h\partial_x b,\\
&\partial_t(hv) + \partial_x(huv) + \frac{\partial}{\partial y}\left(huv + \frac{h^2}{2}\right) = -h\partial_y b,
\end{aligned}$$

and can be condensed, in dimensionalized version, as the following form :

$$\begin{aligned}
&\partial_t h + \operatorname{div}(h\mathbf{u}) = 0,\\
&\partial_t(h\mathbf{u}) + \operatorname{div}(h\mathbf{u}\otimes\mathbf{u}) + \boldsymbol{\nabla}\left(\frac{gh^2}{2}\right) = -gh\boldsymbol{\nabla}b.
\end{aligned}$$

This is the *Shallow-Water* system.

Bibliography

[1] A. Duran and K. Saleh. *Analyse théorique et numérique des équations de Saint-Venant*. Institut Camille Jordan, 2022.

[2] A. Haidar, F. Marche, and F. Vilar. A posteriori finite-volume local subcell correction of high-order discontinuous Galerkin schemes for the nonlinear shallow-water equations. *J. Comput. Phys.*, 452:110902, 2022.

[3] A. Haidar, F. Marche, and F. Vilar. A robust dg-ale formulation for nonlinear shallow-water interactions with a floating object - part ii: Moving object. *J. Comput. Phys.*, under revision, 2022.

[4] F. Vilar. A posteriori correction of high-order discontinuous galerkin scheme through subcell finite volume formulation and flux reconstruction. *J. Comput. Phys.*, 387:245–279, 2019.