UNIVERSITÀ
DEGLI STUDI
DI PALERMO

PhD joint program: University of Catania - University of Messina
XXXIV cycle

Doctoral Thesis

Shock-Capturing methods: Well-Balanced Approximate Taylor and Semi-Implicit schemes

Author:
Emanuele Macca

Supervisor:
Prof. Giovanni Russo

Head of the Doctoral School:
Prof.ssa Maria Carmela Lombardo

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Mathematics and Computational Sciences

Science is the belief

in the ignorance of experts

(R. Feynman)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In nature, economics, and in science in general, a large number of events can be described by differential equations of various types. There are many categories of differential equations based on the type of interaction between the subjects under examination. We pass from the ordinary differential equations, used for example in the evolutionary process of a population, to the partial differential equations, used in phenomena that examine the interactions between different variables, such as interactions between populations or a pollutant in a river.

A partial differential equation (PDE) is an equation involving an unknown function of two or more variables and some of its partial derivatives.

For the sake of simplicity let us introduce some notations.

- $u = u(x_1, \dots, x_s) : \Omega \to \mathbb{R}$ is a $s-$variables function where $\Omega \subseteq \mathbb{R}^s$.
- Let be $\alpha = \alpha_1 + \cdots + \alpha_s$ and k a nonnegative integer, a $k-th$ order derivative of u is so defined
$$D^k u(x) := \frac{\partial^\alpha u(x)}{\partial x_1^{\alpha_1} \dots \partial x_s^{\alpha_s}}.$$

Definition 1.0.1 *Given $\Omega \subseteq \mathbb{R}^s$ and an unknown function $u : \Omega \to \mathbb{R}$, a scalar $k-th$ order partial differential equation is an expression of the form*

$$\mathcal{H}\Big(D^k(u(x)), D^{k-1}(u(x)), \dots, D(u(x)), u(x), x\Big) = 0 \tag{1.0.1}$$

where $\mathcal{H}$ is a given function $\mathcal{H} : \mathbb{R}^{s^k} \times \mathbb{R}^{s^{k-1}} \times \dots \mathbb{R}^s \times \mathbb{R} \times \Omega \to \mathbb{R}$ and $D^k(u(x))$ is a $k-th$ derivatives of u.

Find the solutions of a partial differential equation means determine the set of functions u that verify the condition (1.0.1). Often, these functions are not uniquely defined but vary according to different boundary conditions imposed on $\partial\Omega$. In general, there is no unique analytical methodology to solve any partial differential equation. However, a number of analytical and numerical techniques have been developed to find one or more solutions according to the boundary conditions [12, 29, 122, 123]. In this regard, it is appropriate to distinguish the partial differential equations into 4 typology.

Definition 1.0.2 *The partial differential equations could be differentiate into 4 families of PDE: linear, semi-linear, quasi-linear and fully non-linear.*

- *A partial differential equation is said Linear if it is linear in the unknown function and in all derivatives with coefficients dependent on independent variables. So if it has the form*

$$\sum_{\ell=0}^{k} a_\ell(x) D^\ell(u(x)) = f(x),$$

where $D^0(u(x))$ represents $u(x)$. Notice that $\ell = 0, \ldots, k$ in our case represents $\sum_{i=1}^{s} l_i = \ell \in \{0, \ldots, h\}$.
If $f \equiv 0$ the linear PDE is said homogeneous.

- *A PDE is said semi-linear if it is linear in the maximum order derivatives with dependent coefficients depending on the independent variables. It has the form*

$$\sum_{\ell=k} a_\ell(x) D^\ell u + a_0(D^{k-1}u, \ldots, Du, u, x\Big) = 0.$$

- *A PDE is said quasi-linear if it is linear in the maximum order derivatives with dependent coefficients depend on the independent variables, the unknown function and the derivatives of the function unknown. It has the form*

$$\sum_{\ell=k} a_\ell(D^{k-1}u, \ldots, Du, u, x\Big) D^\ell u + a_0(D^{k-1}u, \ldots, Du, u, x\Big) = 0.$$

- *A PDE is said fully non-linear if it depends non-linearly upon the highest order derivatives.*

A system of partial differential equations is a collection of several partial differential

equation for different unknown functions but for the moment we will focus on a single partial differential equation.

1.1 Scalar conservation laws

For sake of simplicity, let us consider the one-dimensional non-linear scalar conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{1.1.1}$$

where $u = u(x,t) : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$, $u(x,0) = u_0(x)$, and $f(u) = f(u(x,t))$. Under enough differentiable hypothesis for u and $f(u)$, equation (1.1.1) can be expressed as

$$\frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = 0,$$

that, in literature, is named *wave problem* with a wave speed $f'(u)$. The characteristics $x(t)$ are given by

$$\frac{dx(t)}{dt} = f'(u),$$

with initial conditions $x(0) = \xi_0 \in \mathbb{R}$. Known that, the solution is constant along the characteristics, i.e. $u(x,t) = u_0(x - f'(u)t)$, and applying the Dini Theorem [31, 32], we find

$$\frac{\partial u}{\partial x} = -\frac{u_0'}{1 + u_0' f''(u)t} \quad \text{and} \quad \frac{\partial u}{\partial t} = \frac{u_0' f'(u)}{1 + u_0' f''(u)t}.$$

Hence, in cases in which $u_0' f''(x)$ is negative ($u_0' f''(x) < 0$), the derivatives become unbounded with t increasing. In particular, if $f(u)$ is a convex function ($f''(u) > 0$), any initial condition with a negative gradient ($u_0' < 0$), will result in the formation of a discontinuity in finite time. The formation of the previous discontinuity is due to the crossing of the characteristics related to different initial conditions ξ_0. Fortunately, once the discontinuity is formed and propagates, characteristics may again make sense and the discontinuity itself will propagate along a characteristic.

From an analytic point of view, let us suppose that the discontinuity is located at position $x = 0$, where for simplicity $x \in [-L, L]$, and denoting u_l and u_r as left and right state of the

solution, we obtain

$$\frac{d}{dt}\int_{-L}^{L} u(x,t)dx = \int_{-L}^{L} \frac{\partial}{\partial t}u(x,t)dx = \int_{-L}^{L} -\frac{\partial}{\partial x}f(u(x,t))dx = f(u_l) - f(u_r).$$

Furthermore, if we assume that the discontinuity propagates at the constant speed s, conservation of mass requires

$$\int_{-L}^{L} u(x,t)dx = (L+st)u_l - (L-st)u_r,$$

from which we recover

$$\frac{d}{dt}\int_{-L}^{L} u(x,t)dx = s(u_l - u_r) = f(u_l) - f(u_r).$$

As consequence of this phenomena, the following condition is satysfied

$$s[u] = [f], \quad \text{in which} \quad [v] = v_l - v_r,$$

also known as the Rankine-Hugoniot (R-H) jump condition [64, 101]. From the R-H conditions we obtain the speed of propagation for the discontinuity as

$$s = \frac{[f]}{[u]} := \frac{f(u_l) - f(u_r)}{u_l - u_r}.$$

Since a discontinuity is formed, the meaning of classical derivative and consequently classical solution is loss and a new family of solutions, named weak solutions, may be considered.

1.1.1 Weak solutions

Let us consider the scalar non-linear conservation law and in particular the integral form over the domain $[a,b]\times[t_1,t_2]$

$$\int_a^b u(x,t_2)dx - \int_a^b u(x,t_1)dx = \int_{t_1}^{t_2} (f(x_1,t) - f(x_2,t))dt. \tag{1.1.2}$$

Definition 1.1.1 *A locally integrable function u is a weak solution of* (1.1.1) *if*

$$\int_a^b \int_{t_1}^{t_2} [u_t + f(u)_x]\phi(x,t)dxdt = 0, \tag{1.1.3}$$

for every differentiable function with compact support $\phi \in \mathcal{C}_0^1([a,b]\times[t_1,t_2])$.

Integrating by parts we follow that

$$\int_a^b \int_{t_1}^{t_2} [U\phi_t + f(U)\phi_x]dxdt = -\int_a^b u(x,0)\phi(x,0)dx. \tag{1.1.4}$$

Weak solutions are an appropriate generalization of the classical solutions for systems of conservation and balance laws. It could be proved that classical solutions are also weak solutions and, if the solution is continuously differentiable, then weak solution implies classical one [11, 38, 40, 107]. However, the connection between solution (1.1.2) and (1.1.4) is automatic and we will refer to $u(x,t)$ as a weak solution to the conservation law (1.1.1) provided that satisfies (1.1.3) for all admissible test functions.

A numerical version of the Rankine-Hugoniot conditions [64, 101] can be found in [25, 58, 59]. These conditions are able to characterize the weak solution in terms of jumps and discontinuity movements giving information about the behaviour of the conserved variables close to the discontinuities.

With the introduction of this new family solution, i.e. *weak solution*, we are able to transform the classical derivatives into the weak derivatives throughout the test functions $\phi(x,t)$. However, this approach leads the uniqueness of the solution making necessary the introduction of new hypotheses/conditions in order to obtain uniqueness.

1.1.2 Entropy conditions

To overcome the issue of multiple weak solutions, we need a criteria to determinate, if exists, whether a solution is admissible. With this in mind let us introduce the Entropy condition to guarantee the uniqueness of weak solution.

A natural condition of admissibility, called entropy condition [81], is given by

$$\frac{f(u)-f(u_r)}{u-u_r} \leq s \leq \frac{f(u)-f(u_l)}{u-u_l} \quad \forall u \in (u_l,u_r), \tag{1.1.5}$$

reflecting that the characteristics must run into the shock.

Condition (1.1.5) is called Lax's E-condition [80]. There is also an entropy inequality for entropy-entropy flux pairs, due also to Lax [80], which is closely linked to vanishing viscosity solutions. There are other type of entropy conditions as Oleinik's E-condition [95], Kruzkov's condition [73], Wendroff's condition [126] or Liu's condition [87].

Theorem 1.1.1 *Let u and v be piecewise smooth weak solutions to* (1.1.1) *with a convex flux and assume that all discontinuities are shocks. Then*

$$\frac{d}{dt}\|u(t) - v(t)\|_1 \leq 0.$$

The property

$$\frac{d}{dt}\|u(t) - v(t)\|_1 \leq 0$$

usually known as L^1-contraction involves a list of immediate consequences as:

- If u is a weak solution to (1.1.1) with a convex flux and it satisfies an entropy condition, the solution is unique.

- If a discontinuity violates the entropy condition, then there is a solution, v, such that

$$\frac{d}{dt}\|u(t) - v(t)\|_1 \geq 0.$$

Even if, the combination of entropy condition and the Rankine-Hugoniot condition ensure the uniqueness of the weak solution it may be too restrictive for a general initial condition $u_0(x)$.
With this in mind following [57, 81], let us introduce the vanishing viscosity solution for the conservation law (1.1.1) as the solution of the modified equation

$$\frac{\partial u_\varepsilon}{\partial t} + \frac{\partial f(u_\varepsilon)}{\partial x} = \varepsilon \frac{\partial^2 u_\varepsilon}{\partial x^2} \tag{1.1.6}$$

with initial condition $u_\varepsilon(x, 0) = u_0(x)$.

Theorem 1.1.2 *Let be $f(u)$ a convex flux. In the limit of $\varepsilon \to 0$, the vanishing solution of* (1.1.6) *is a weak solution of* (1.1.1) *and satisfies the entropy condition.*

The key results in much of above discussion rely on the assumption of a convex flux, thus limiting their validity to a scalar conservation laws in one-dimension. However, all the results could be extended to more general entropy condition for multiple-dimensions [113] and systems of conservation laws redefining the entropy condition appropriately for systems of conservation and balance laws [6, 11, 30, 38, 47, 57, 60, 61, 76, 80, 82, 93, 113].

1.2 Numerical Framework

Given a differential equation, in our case an hyperbolic conservation and balance law, a large number of different methods could be developed to solve numerically the equation following several different approaches based one the reconstruction strategy as: finite difference, finite volume etc. In this elaborate we focus on finite difference and finite volume numerical methods.

1.2.1 Finite Difference

Finite difference method techniques are based on approximations that allow the substitution of differential equations in finite difference equations. These finite differences are generally approximate in algebraic form; they relate the value of the dependent variable at a point to the values of the neighboring points. Thus a finite differences solution involves several stages:

- Divide the solution domain into a node grid, i.e. discretize the domain.
- At each node of the grid, the differential equation is approximate by replacing partial derivatives with appropriate approximations, obtained in terms of values at the unknown function, i.e. approximate the differential equation given in the equivalent finite difference equation by connecting the unknown function at a point with the values of its neighboring points.
- The result is an algebraic equation for each node, which contains the unknown in the node itself and in some adjacent nodes, adding prescribed boundary conditions and/or initial conditions.

The finite-difference method is defined dimension per dimension; this makes it easy to increase the "element order" to get higher-order accuracy. If the simulation can fit in a rectangular or box-shaped geometry using a regular grid, efficient implementations are much easier than for finite-volume and other types of methods. Regular grids are useful for very-large-scale simulations on supercomputers often used in meteorological, seismological, and astrophysical simulations.
The finite-difference method may run into problems handling curved boundaries for the purpose of defining the boundary conditions. Boundary conditions are needed to truncate

the computational domain.
For computations that need high accuracy, the extra effort in making boundary-fitted meshes and the associated complications of such meshes for the implementation may be worth it.

Such methods are arguably the first family of methods that have been proposed to solve differential equations. The finite different approximation were known to Euler and the first analysis of finite difference methods for initial value problems was presented by Courant, Friedrichs and Lewy on 1928 [28] in which the CFL condition was introduced.

Crossing to numerical point of view, let us define the grid operators Δ^+ and Δ^- as

$$\Delta^+ f = f_1 - f_0 \quad \Delta^- f = f_0 - f_{-1},$$

and the difference operators D_h^+, D_h^0 and D_h^- as

$$D_h^+ = \frac{\Delta^+}{h} \quad D_h^- = \frac{\Delta^-}{h} \quad D_h^0 = \frac{\Delta^+ + \Delta^-}{2h},$$

where $h = \frac{b-a}{N}$ is the step grid based on the spatial grid $x_i = a + ih$ for $i = 0, \ldots, N$, defined on the interval $[a, b]$. Observe that the interval $[a, b]$ is discretize in $N+1$ points where $x_0 = a$ and $x_N = b$. In a similar way, we have the temporal grid $t_n = t_0 + nk \in [t_0, t_f]$, where t_0, t_f are, respectively, the initial and final time and k is the step time, usually related to the space step and CFL condition.
The difference operators can be applied in both space and time derivative to represent the spatial and temporal derivatives.
The space and time derivatives may be approximated by the difference operators through the Taylor expansion. Indeed,

$$D_h^{\pm} u(x_i) = u'(x_i) \pm \frac{h}{2} u''(x_i) + \mathcal{O}(h^2) \quad D_h^0 u(x_i) = u'(x_i) \pm \frac{h^2}{6} u^{(3)}(x_i) + \mathcal{O}(h^4)$$

Let us consider the scalar conservation law (1.1.1) and for simplicity periodic boundary conditions. We define the space-time grid as the couple $(x_i, t_n) = (a + ih, t_0 + nk)$, and the numerical scheme

$$u_i^{n+1} = G(u_{i-q}^n, \ldots, u_{i+p}^n)$$

where $u_i^n \approx u(x_i, t_n)$.

As numerical approximation, the idea under the finite difference methods is to generate

a sequence of grid gradually finer that are an approximation of the exact solution. For this reason, the main problem is to check consistency, stability and convergence of these sequence through the analysis of local error.

For sake of simplicity, let us consider linear scalar conservation law and a 3-points numerical scheme

$$u_i^{n+1} = G(u_{i-1}^n, u_i^n, u_{i+1}^n).$$

Analytically,

$$u(x_i, t_{n+1}) = G(u^n)_i + k\tau_i^n$$

where τ_i^n is the truncation error defined as the difference between the exact solution and the numerical one. Defining the local error as $\varepsilon_i^{n+1} = u(x_i, t_{n+1}) - G(u^n)_i$, we have the related initial condition error $\varepsilon_i^1 = G(\varepsilon^0)_i + k\tau_i^0$, under the hypothesis the G is linear. In general,

$$\varepsilon_i^n = G^n(\varepsilon^0)_i + k\sum_{j=0}^{n-1} G^{n-j-1}(\tau^j)_i$$

where G^n is the composition of n-time G.

Let be $\|\cdot\|_{h,q}$ the q-norm at discrete level with step h, we observe that

$$\begin{aligned}\|\varepsilon^n\|_{h,q} &\leq \|G^n(\varepsilon^0)\|_{h,q} + k\sum_{j=0}^{n-1}\|G^{n-j-1}(\tau^j)_i\|_{h,q} \\ &\leq \|G^n\|_{h,q}\|\varepsilon^0\|_{h,q} + kn\max_j \|G^{n-j-1}\|_{h,q}\|\tau^j\|_{h,q}\end{aligned}$$

where we are adopting the standard definition of the subordinate matrix norm.

We will say that the scheme is *consistent* if

$$\lim_{l\to\infty}\begin{cases}\|\varepsilon^0\|_{h_l,q} \\ \max_j \|\tau^j\|_{h_l,q}\end{cases} = 0.$$

Furthermore, we say that a scheme is of order (s,t) if $\|\tau\|_{h,q} = \mathcal{O}(h^s, k^t)$.

A numerical methods is *stable* if there exist $C > 0$ such that

$$\|G^n\|_{h,q} \leq C$$

definitively in time [75, 103, 115].

1.2.2 Finite Volume

The finite volume method is a discretization technique for partial differential equations, especially those that arise from physical conservation laws. The finite volume method is a method for representing and evaluating partial differential equations in the form of algebraic equations [83].

The finite volume method makes use of the integral form of conservation equations and also develops in several phases:

- The domain is divided into control volumes called also cells, and conservation equations are applied to each volume.

- Usually the variable lies in the centre of the cell.

- Interpolations are used to express the values of the variables, or of the gradients, on the cell surfaces, and it is necessary to approximate surface (flow) and volume integrals.

- As a result, an algebraic equation is obtained (by analogy with finite difference) for each cell, and then a system of equations.

In the finite volume method, volume integrals in a partial differential equation that contain a divergence term are converted to surface integrals, using the Divergence Theorem [46, 51, 104]. These terms are then evaluated as fluxes at the surfaces of each finite volume. Because the flux entering a given volume is identical to that leaving the adjacent volume, these methods are conservative by definition. Another advantage of the finite volume method is that it is easily formulated to allow for unstructured meshes.
The finite-volume method's strength is that it only needs to do flux evaluation for the cell boundaries. This also holds for nonlinear problems, which makes it extra powerful for robust handling of (nonlinear) conservation laws appearing in transport problems.

The local accuracy of the finite-volume method, such as close to a corner of interest, can be increased by refining the mesh around that corner. However, the functions that approximate the solution when using the finite-volume method cannot be easily made of higher order. This is a disadvantage of the finite-volume methods compared to the finite-difference ones.

Let us consider again the space-time gride $(x_i, t_n) = (hi, kn)$, where h and k are respec-

tively the space and time step size. Let us consider also the space-time cell so defined

$$I_i^n = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t_n, t_{n+1}] \quad \text{where} \quad x_{i\pm\frac{1}{2}} = x_i \pm \frac{1}{2}h.$$

Integrating the scalar conservation laws (1.1.1) on the space-time cell I_i^n we obtain

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \Big[u(x, t_{n+1}) - u(x, t_n)\Big] dx = \int_{t_n}^{t_{n+1}} \Big[f(u(x_{i-\frac{1}{2}}, t)) - f(u(x_{i+\frac{1}{2}}, t))\Big] dt.$$

Let be, $\bar{u}_i^n$, the space cell average at time t_n so defined

$$\bar{u}_i^n = \frac{1}{h} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_n) dx,$$

and $F_{i+\frac{1}{2}}^n$ the time cell average in position $x_{i+\frac{1}{2}}$ as

$$F_{i+\frac{1}{2}}^n = \frac{1}{k} \int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}}, t)) dt. \tag{1.2.1}$$

A finite volume method is written in the following form

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{k}{h}(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n).$$

The problem is then how to compute the flux at cell boundaries.
Differently from the finite difference method, the aim is previously to compute the approximation of u at boundary cell, i.e. in position $x_{i+\frac{1}{2}}$ at time t_n, using the stencil points $\bar{u}_{i-q}^n, \ldots, \bar{u}_{i+p}^n$. Reconstructed $u_{i+\frac{1}{2}}^*$ as $u_{i+\frac{1}{2}}^* = \mathcal{R}(\bar{u}_{i-q}^n, \ldots, \bar{u}_{i+p}^n)$ where $\mathcal{R}$ is usually a polynomial reconstruction $p(x)$ that satisfies theoretically the conservation

$$\int_{x_{l-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p(x) dx = \bar{u}_i^n \quad \forall j \in \{i-q, \ldots, i+p\},$$

the flux at cell boundaries is so computed $F_{i+\frac{1}{2}}^n \approx f(u_{i+\frac{1}{2}}^*)$

Remark 1.2.1 *Finite volume methods can be compared and contrasted with the finite difference methods, which approximate derivatives using nodal values and construct a global approximation by stitching them together. In contrast a finite volume method evaluates exact expressions for the average value of the solution over some volume, and uses this data to*

construct approximations of the solution within cells.

1.2.3 Numerical properties

As we have said before, the simplest way to approximate derivatives is by means of linear finite differences. Unfortunately sometimes, finite differences methods do not yield a satisfactory approximation of the partial derivative appearing in the equations when a singularity solution is considered. Finite volume methods overcome this difficulty by resorting to weak formulation that do not require derivatives of the unknowns.

However, there exists a simple requirement that we can impose on the numerical methods to guarantee that they do not converge to non-solutions. The Lax-Wendroff's theorem guarantee the convergence to the theoretical solution for the conservative schemes.

Definition 1.2.1 *Let us consider a non-linear scalar equation of balance law written in the following form*

$$u_t + f(u)_x = S.$$

A numerical method is said to be conservative if it can be written in the form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}) + \Delta t S_i^n, \tag{1.2.2}$$

where

$$F_{i-\frac{1}{2}} = \mathcal{R}(u_{i-p}^n, \ldots, u_{i+q-1}) \quad \text{and} \quad F_{i+\frac{1}{2}} = \mathcal{R}(u_{i-p+1}^n, \ldots, u_{i+q})$$

in which $p, q \in \mathbb{N}$ with $p, q \geq 0$. $\mathcal{R}$ in called numerical flux function, that for finite difference methods is included in operator G.

As said before, the finite volume schemes are conservative by construction, however the aim of the conservative schemes is to reproduce exactly, at discrete level, the conservation of the physical quantities.

Definition 1.2.2 *A numerical flux function of a conservative numerical method is consistent with the conservation or balance laws if the numerical flux function $\mathcal{R}$ reduces to the exact flux f when applied to constant flow, i.e.*

$$\mathcal{R}(u, \ldots, u) = f(u).$$

The Lax-Wendroff's theorem proves that, given a conservative method that produces a sequence of approximations that converges to some function $u(x,t)$ as the grid is refined, then this function will be a weak solution of the conservation or balance law [61, 76].

Theorem 1.2.1 *Consider a sequence of grid indexed by $\ell = 1, 2, \ldots$ with grid sizes respectively Δx_ℓ and Δt_ℓ satisfying*

$$\lim_{\ell \to \infty} \Delta x_\ell = 0,$$

$$\lim_{\ell \to \infty} \Delta t_\ell = 0.$$

Let $\{u_\ell(x,t)\}$ denote the piecewise constant function obtained by a conservative methods on the $\ell-$th grid. If the total variation of the function $u_\ell(\cdot,t)$ is uniformly bounded in ℓ and t, i.e. $\sup_{\ell,t\in[0,T]} TV(u_\ell(\cdot,t)) < \infty$ and $u_\ell(x,t)$ converge in L^1_{loc} to a function $u(x,t)$ as $\ell \to \infty$, then u is a weak solution of the conservation or balance law.

1.2.4 High-resolution conservative methods

The term high-resolution is generally applied to methods whose the local truncation error has order higher than two. This strategy allows to obtain a numerical solution that has a second or even higher order global errors in the region in which the solution is smooth; while returns well-resolved non-oscillatory approximations near discontinuities.

Although, there are several high-order resolution techniques for conservative methods. We will focus on the Approximate Taylor approach based on the Lax-Wendroff methods which, unlike linear methods, reconstruct solutions in time and space at the same time. Nevertheless, the method of lines, which refers to numerical methods for PDEs that, first discretizes the spatial derivatives only and leaves the variables of continuous time. This lead to a system of ordinary differential equation to which a numerical method can be applied for ordinary equations of initial value. In chapter 3, we use for spatial reconstruction the well-known scheme called Weighted Essentially Non-Oscillatory (WENO) methods [67] or [88] in a conservative form. And for discretization in time, the high order TVD Runge-Kutta methods is applied.

Finite difference WENO

The Essentially Non-Oscillatory reconstruction (ENO), for a given cell interface reconstruction, is obtained by choosing one of the different polynomial reconstructions of a fixed degree

that can be constructed using stencils that contain one of the cells that define the given interface. The stencil choice is based on the smoothness of the numerical solution defined on the stencil and the obtained reconstructions are $r-th$ order accurate when considering r stencils (consecutive indexes) of length r containing the target cell, with the condition that at least one of the stencils does not contain a singularity.

During the stencil selection procedure, the ENO method considers r possible stencils, which in total contain $2r-1$ cells.

WENO reconstructions, introduced by Liu, Osher and Chan in [88], are based on the idea of increasing the order of accuracy of the method in smooth regions by considering a reconstruction found by a convex combination of the different polynomial reconstruction candidates of the ENO scheme, with spatially varying weights designed to increase the accuracy of the individual reconstructions corresponding to different stencils.

For sake of simplicity, since the WENO reconstructions have been adopted just for systems of conservation laws (Chapter 3), let us consider the scalar conservation law

$$u_t + f(u)_x = 0.$$

Let consider a uniform mesh of constant step $h = \Delta x$ and nodes $\{x_i\}$ and the following notation will be used for the intercells

$$x_{i+\frac{1}{2}} = x_i + \frac{1}{2}h, \quad \forall i = 1, \dots, N,$$

where N represents the number of nodes.
The methods presented here was developed by Shu and Osher in [112]. In their approach, the discretization of the derivative of the flux follows as:

$$f(u(x,t))_x = \frac{\hat{f}(x+\frac{1}{2}h) - \hat{f}(x-\frac{1}{2}h)}{h}$$

that is exactly satisfied if $\hat{f}(x)$ is a function such that

$$\frac{1}{h}\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} \hat{f}(\sigma)d\sigma = f(u(x,t)).$$

With this in mind, the semi-discrete method can be written in conservative form as follows:

$$\frac{du_i}{dt} + \frac{1}{h}\Big(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}\Big) = 0, \tag{1.2.3}$$

where

$$\hat{f}_{i+\frac{1}{2}} = \mathcal{R}(f(u_{i-r+1}), \dots, f(u_{i+r})) \tag{1.2.4}$$

is an approximation of $\hat{f}(x_{i+\frac{1}{2}})$. Once, the procedure has been defined, let us focus on the WENO reconstruction operator $\mathcal{R}$. For this reason, let us consider the family of r possible stencils containing r nodes

$$S^r_{i+\ell} = \{x_{i+\ell-r+1}, \dots, x_{i+\ell}\}$$

with $\ell = 0, \dots, r-1$. From them, following the Essentially Non-Oscillatory (ENO) procedure, r different polynomial reconstructions of degree at least $r-1$, $p^r_\ell(x)$, can be constructed such that satisfy

$$p^r_\ell(x_{i+\frac{1}{2}}) = f(u(x_{i+\frac{1}{2}}, t)) + \mathcal{O}(h^r)$$

for all time t and under sufficient regularity of f. Liu et al. in [88] provide that there is no need of selecting just one of the possible stencils and, considering a combination of them, better results could be obtained in smooth region. For this reason, a $(2r-1)-th$ order reconstruction

$$p^{2r-1}_{r-1}(x_{i+\frac{1}{2}}) = f(u(x_{i+\frac{1}{2}}, t)) + \mathcal{O}(h^{2r-1})$$

can be computed on stencil

$$S^{2r-1}_{i+r-1} = \{x_{i-r+1}, \dots, x_{i+r-1}\},$$

instead of the $r-th$ order reconstruction obtain with the ENO procedure.

Considering the r candidate stencils of the ENO procedure, $S^r_{i+\ell} = \{x_{i+\ell-r+1}, \dots, x_{i+\ell}\}$, with $\ell = 0, \dots, r-1$, and the $(r-1)-th$ degree polynomial reconstructions $p^r_\ell(x)$, related to the stencil $S^r_{i+\ell}$, that satisfy

$$p^r_\ell(x_{i+\frac{1}{2}}) = f(u(x_{i+\frac{1}{2}}, t)) + \mathcal{O}(h^r),$$

the WENO reconstruction of f is fixed as the convex combination between the $(r-1)-th$

polynomials as:

$$\hat{f}_{i+\frac{1}{2}} = \mathcal{R}(f(u_{i-r+1}), \ldots, f(u_{i+r-1})) = \sum_{\ell=0}^{r-1} \omega_\ell p_\ell^r(x_{i+\frac{1}{2}}), \tag{1.2.5}$$

where the weight are defined as:

$$\omega_\ell = \frac{\alpha_\ell}{\sum_{j=0}^{r-1} \alpha_j}, \quad \alpha_\ell = \frac{C_\ell^r}{(\varepsilon + I_\ell)^p}$$

for all $\ell = 0, \ldots, r-1$, where $p \in \mathbb{N}$; C_ℓ^r are the optimal weight, i.e. the positive coefficients such that

$$p_{r-1}^{2r-1}(x_{i+\frac{1}{2}}) = \sum_{\ell=0}^{r-1} C_\ell^r p_\ell^r(x_{i+\frac{1}{2}}) \quad \text{and} \quad \sum_{\ell=0}^{r-1} C_\ell^r = 1;$$

$I_\ell = I_\ell(h)$ is an smoothness indicator of f in stencil S_ℓ defined as

$$I_\ell = \sum_{q=1}^{r-1} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} h^{2q-1} \left(p_\ell^{(q)}(x) \right)^2 dx;$$

and ε is a small positive number to avoid null denominators.

High order TVD Runge-Kutta scheme

To achieve high order accuracy in time discretizations , one can use the Total Variation Diminishing (TVD) Runge-kutta method, due they ensure that the total variation of the solutions does not increase (under some time step restrictions) [49, 109].

The methods solve the semi-discretized system of ordinary differential equations ODEs

$$u_t = L(u),$$

with a suitable initial conditions, resulting from a methods of lines approximation to a hyperbolic conservation law:

$$u_t = -f(u)_x,$$

where, $-f(u)_x$ is approximate by some types of spatial discretizations. For this work, we

consider the 3ª order 3 stages TVD Runge-Kutta method defined as:

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

with effective CFL= 0.33, which is known as the Shu-Osher method [111]. And the 4ª order method with 10 stages and effective CFL= 0.6,

$$\begin{aligned} u^{(1)} &= u^n + \frac{1}{6}\Delta t f(u^n) \\ u^{(i+1)} &= u^{(i)} + \frac{1}{6}\Delta t f(u^{(i)}), \quad i = 1, 2, 3 \\ u^{(5)} &= \frac{3}{5}u^n + \frac{2}{5}(u^{(4)} + \frac{1}{6}\Delta t f(u^{(4)})) \\ u^{(i+1)} &= u^{(i)} + \frac{1}{6}\Delta t f(u^{(i)}), \quad i = 5, 6, 7, 8 \\ u^{n+1} &= \frac{1}{25}u^n + \frac{9}{25}(u^{(4)} + \frac{1}{6}\Delta t f(u^{(4)})) + \frac{3}{5}(u^{(9)} + \frac{1}{6}\Delta t f(u^{(9)})). \end{aligned}$$

This method belongs to the family of Strong Stability Preserving Runge-Kutta schemes (SSPRK) see [48].

1.3 Thesis Framework

This section, entirely discursive, outlines the motivations, the objectives and the structure in which, the thesis, has been written.

1.3.1 Motivation

Lax-Wendroff type schemes for linear systems of conservation laws by construction are strictly related to the Taylor expansions in time in which the time derivatives are replaced by the spatial derivatives through the governing equations [76, 84, 120]. The spatial derivatives are successfully discretized by means of centered high-order differentiation formulas. This strategy allows to derive numerical methods of order R, where a selected centered $(R+1)-$point stencil must be used. In this case, the schemes are L^2-stable under the usual Courant-Friedrichs-Lewy (CFL) condition [28]. This thesis is mainly focused on the

extension of Lax-Wendroff type methods to nonlinear systems of conservation and balance laws emphasizing the well-balanced methods and deals with a *work in progress* developing an IMEX strategy for the Exner model of shallow water with sedimentation.

About the first topic, LW-type schemes have already been considered in the literature, as a possible alternative to multistep or multistage one-step methods such as the original finite volume ENO schemes (see [56]) and the SSP Runge-Kutta schemes (see [49]). LW-type approach was followed by E.F. Toro and collaborators in the design of the so-called ADER (arbitrary high-order schemes utilizing higher order derivatives) methods: see [108, 118, 121]. The computation of time derivatives in these methods is based on the modified generalized Riemann problem introduced by Toro in [119, 120]. A Lax-Wendroff, second order evolution, Galerkin method for multidimensional hyperbolic systems was also introduced in [90]. More recently (2003), in [99] this procedure has been used together with WENO reconstructions for the spatial discretization. The main benefit, compared to RK time discretizations, is that only one WENO reconstruction is needed at each spatial cell per time step. The main difficulty to extend Lax-Wendroff methods to nonlinear problems comes from the transformation of time derivatives into spatial derivatives through the governing equations. A first strategy to do this was given by the Cauchy-Kovalevskaya (CK) procedure, in which the PDE is used to replace time derivatives by spatial derivatives. The main drawback of this procedure comes from the increasing computational cost when an high order scheme is considered. In the context of ADER methods, this difficulty was accurately circumvented in ADER-WENO methods (see [37]) by replacing the CK procedure by local space-time problems that are solved with a Galerkin method. The so-called PNPM methods introduced in [35], that generalize ADER-WENO and DG methods, also follow this approach. These methods can be applied both on structured and unstructured meshes with CFL-1 condition for stability.

An alternative to both CK and local space-time problems has been recently proposed in [132] based on an Approximate Taylor (AT) method: the time derivatives are approximated using high-order centered differentiation formulas combined with Taylor approximations in time that are computed in a recursive way. Unfortunately, AT schemes when applied to linear systems do not recover classical Lax-Wendroff methods: indeed, they have $(2R+1)-$point stencils and worse linear stability properties than the original $R-$order ($(R+1)-$stencil) Lax-Wendroff methods. Nevertheless, they can be stabilized by using one WENO reconstruction per spatial cell and time step, as shown in [99] even for linear problem, and the resulting

methods typically give good results under a CFL≤ 0.5 condition. In 2019, Carrillo and Parés, developed a compact version of the AT schemes (CAT) that is a properly generalization of the Lax-Wendroff methods (see Chapter 3) [15]. As it is well known, Lax-Wendroff-type reconstructions produce spurious oscillations when a discontinuity appears at discrete level [55]. Several strategies have been developed to reduce and damp these oscillations: flux limiters [70, 116]; essentially non-oscillatory reconstructions like ENO [56] or WENO [111, 112] or CWENO [85, 86]; MOOD approach [26]; order-adaptive approach [14].

About the second topic, the focus is to introduce an IMEX strategy to compute the sediment evolution [9, 18] in the Exner model of sediment transport in shallow water in order to improve improve both stability and efficiency. The Exner model is a system of PDEs that coupled the shallow water equations with a tranport equation for the sediment, in this work the Grass equation [50]. After some manipulation, the Exner model can be written as a non-conservative hyperbolic system

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = 0.$$

This system is strictly hyperbolic if and only if the characteristic polynomial has three distinct real roots $\lambda_1 < \lambda_2 < \lambda_3$. Under the hypothesis of Froude number (Fr) less then 1 it is $\lambda_1 < 0$ and $\lambda_3 > 0$. Assuming that the interaction between the water and the sediment is weak, it is $\lambda_2 \leq \min(|\lambda_1|, |\lambda_3|)$, i.e. the wave speed of the sediment is much smaller than the water wave speeds. An explicit method implies a strong stability restriction due to the velocity of the free-surface wave. This restriction involves in a very long computation time that could be reduced neglecting the behaviour of the free-surface waves behaviour and looking at the sediment evolution. The objective is to drastically improve the efficiency in the computation of the evolution of the sediment by treating water waves implicitly, thus allowing much larger time steps than the one allowed by standard CFL condition on explicit schemes. Recently, Garres-Díaz et al. (2022) proposed a semi-implicit $\theta-$method approach for sediment transport models [44] by which, choosing $\theta > \frac{1}{2}$ in the semi-implicit method, an increase in both efficiency and stability is obtained [22].

1.3.2 Perspectives of the thesis

The thesis has three principally objectives:

1. To develop a limiter approach to reduce and damp the oscillations in the Compact Approximate Taylor schemes for systems of conservation laws.

2. To extend the CAT methods to systems of balance laws emphasizing the non-oscillatory well-balanced schemes.

3. To develop an IMEX strategy for the Exner model of sediment transport, which improves stability and therefore efficiency, over explicit scheme.

1.3.3 Structure of the thesis

These topic have been considered in three papers. The first one, developed in collaboration with Hugo Carrillo and Carlos Parés from University of Málaga (Spain), Giovanni Russo from University of Catania (Italy) and David Zorío from University of Concepcíon (Chile) entitled *An order-adaptive compact approximate taylor method for systems of conservation laws,* has been published on Journal of Computational Physics, pp. 438-31 (2021) [14]. In this article a new family of smoothness indicators has been developed by which a non-oscillatory order-adaptive version of the CAT scheme (ACAT) is presented. The second one, developed in collaboration with Hugo Carrillo and Carlos Parés from University of Málaga (Spain) and Giovanni Russo from University of Catania (Italy) entitled *An order-adaptive compact approximate taylor method for systems of balance laws and relative well-balanced scheme,* has been submitted on Journal of Computational Physics (2002) [13] and is available at the following link https://arxiv.org/abs/2202.02068. This article introduces the extension of the ACAT scheme for systems of balance laws with a particular attention to the high-order well-balanced schemes. The last one, still under development in collaboration with Manuel J. Castro-Díaz from University of Málaga (Spain), Stavros Avgerinos and Giovanni Russo from University of Catania (Italy), introduces a semi-implicit approach for the coupled system shallow water and sediment equations. The aim is reduce the CFL-restrictions due to the free-surface waves in order to increase stability and efficiency.

The thesis is so structured: after a narrative introduction and some preliminary aspects discussed in this chapter, *Chapter 2* contains the Approximate Taylor methods that are the extension of the Lax-Wendroff scheme and the basis from which the CAT schemes have been developed. In particular, firstly the Lax-Wendroff method, for linear and non-linear case, is presented. Successively, the historical extensions of Lax-Wendroff schemes are presented,

such as Mac-Cormack [91] and RitchMeyer [102] method. Then, the Lax-Wendroff-type procedures developed by Qiu and Shu [99] and Zorìo, Baeza and Mulet [132] are presented in detail.

Chapter 3 starts with the general formulation of the high-order Compact Approximate Taylor method for systems of conservation laws. Then, the second-order method is presented, emphasising and comparing the CAT2 scheme with the Lax-Wendroff-type procedures introduced in Chapther 2. Afterwards, the details of CAT2P scheme are presented. Finally, the order-adaptive strategy to avoid the spurious oscillations close the discontinuities for CAT2P has been introduced. Some numerical tests were performed and compared with well-known methods to check the performance of ACAT schemes. At the end, the 2D extension has been presented with some numerical experiments. From this chapter the first paper was born.

Chapter 4 starts with the extension of the high-order Compact Approximate Taylor method for systems of balance laws. Then, the order-adaptive strategy is modified to suit the CAT methods for this extension. Successively, the well-balanced (WBCAT) schemes and its order-adaptive version (WBACAT) are presented. Afterwards, several numerical tests, compared with exact/reference solutions, were performed to check the properties of ACAT and WBACAT schemes for systems of balance laws.

Chapter 5 starts with the extension on the two-dimensional case for the high-order Compact Approximate Taylor method for systems of balance laws. Then, the order-adaptive strategy is presented for this extension. Successively, the two-dimensional well-balanced schemes and its order-adaptive version are presented. Finally, some numerical tests, compared with exact/reference solutions, were performed to check the properties of 2D ACAT and WBACAT schemes for two-dimensional systems of balance laws. From chapters 4-5 the second submitted paper was born.

Chapter 6 starts with a one-dimensional semi-implicit approach for the coupled system between the shallow water equations and the sediment equation. Next, an approximate equation of quasi-stationary states is adopted to evolve the sedimentation in time (see Section 6.3). Some numerical tests to check the behaviour of the schemes are shown. Finally the two-dimensional version with some numerical experiments are considered. The paper on this topic is almost ready for submission.

Chapter 7 shows the thesis conclusions and some future perspectives.

Chapter 2

Approximate Taylor Method

The content of this chapter is designed specifically to introduce the numerical schemes for hyperbolic conservation laws based on the Taylor expansion in time. In particular, our objective is to present the second order Lax Wendroff method applied to linear scalar conservation law and then move to the high order schemes for non-linear systems of conservation laws.

Numerical methods replace the continuous problem represented by the PDEs by a finite set of discrete values. These are obtained, in our case, by first discretising the domain of the PDEs on a mesh. Several discretization techniques are possible such as, for example, finite difference, finite volume, finite element and disconinuous Galerkin discretization. In our thesis we adopt the finite difference discretization in space, for which the discrete values represent a pointwise approximation of the unkown function at grid points. Let consider the *initial boundary value problem* for the linear advection equation in the doman $[a, b] \times [0, T]$ on the Oxt-Cartesian frame.

$$(IBVP) = \begin{cases} \text{PDE}: & u_t + f(u)_x = 0; \\ \text{IC}: & u(x, 0) = u_0(x); \end{cases} \tag{2.0.1}$$

Solving problem (2.0.1) means evolving the solution $u(x, t)$ in time starting from the initial condition $u_0(x)$ at time $t = 0$ and subject to boundary conditions [47].

2.1 Lax Wendroff Method

A scheme of historic as well as practical importance is that of **Lax Wendroff** introduced by Peter Lax and Burton Wendroff in 1960 [59, 76, 77, 79, 120, 125]. It has been the most widely

adopted scheme for aeronautical applications, up to the end of the 1980s under various form.

The original derivation of Lax and Wendroff was based on a Taylor expansion in time up to second order, so as to achieve second order accuracy in time. Thus,

$$u(x_i, t + \Delta t) = u(x_i, t) + \Delta t u_t(x_i, t) + \frac{\Delta t^2}{2} u_{tt}(x_i, t) + O(\Delta t^3) \tag{2.1.1}$$

The scheme is then obtained by neglecting the higher order term in Δt, using the governing equation to replace time derivatives by space derivatives, and then discretising the space derivatives by finite difference approximations.

2.1.1 Linear case

The idea under the Lax-Wendroff scheme is to keep the second time derivative in the discretization and replace all time derivatives by equivalent spatial derivatives through the governing equation (2.0.1). Specifically, in the linear case, the governing equation becomes

$$u_t + a u_x = 0 \tag{2.1.2}$$

and the strategy is straightforward. In fact, for the first time derivative $u_t = -a u_x$; while, the second time derivative is obtained by taking the time derivative of the governing equation (2.1.2). Indeed,

$$u_{tt} = -a(u_x)_t = -a(u_t)_x = a^2 u_{xx}. \tag{2.1.3}$$

Substituting eq (2.1.3) in eq (2.1.1) we obtain

$$u(x_i, t + \Delta t) = u(x_i, t) - a\Delta t u_x(x_i, t) + \frac{a^2 \Delta t^2}{2} u_{xx}(x_i, t) + O(\Delta t^3) \tag{2.1.4}$$

If we discretize all the space derivatives with second order central formulas in the mesh point x_i and neglect higher order terms in Δt and Δx, we get

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{2\Delta x}\Big(u_{i+1}^n - u_{i-1}^n\Big) + \frac{a^2 \Delta t^2}{2\Delta x^2}\Big(u_{i+1}^n - 2u_i^n + u_{i-1}^n\Big), \tag{2.1.5}$$

where $\{x_i\}$ are the nodes of a uniform mesh of step Δx; u_i^n is an approximation of the point value of the solution at position x_i at the time $n\Delta t$, in which Δt is the time step.

A useful alternative formulation of the Lax Wendroff, written in conservative form, is the

following

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\Big(F_{i-\frac{1}{2}}^{\mathrm{LW}} - F_{i+\frac{1}{2}}^{\mathrm{LW}}\Big) \tag{2.1.6}$$

where

$$F_{i+\frac{1}{2}}^{\mathrm{LW}} = \frac{a}{2}\Big(u_{i+1}^n + u_i^n\Big) - \frac{a^2\Delta t}{2\Delta x}\Big(u_{i+1}^n - u_i^n\Big),$$

which emphasizes the conservative structure of the method.

2.1.2 Non-linear case

The Lax-Wendroff scheme for non-linear case gives rise to two families of schemes. The first one contains the natural extension of the Lax-Wendroff one step scheme with the computation of the Jacobian matrix; while the second one contains the Jacobian free schemes.

Jacobian scheme

The derivation of the Lax-Wendroff scheme is not trivial for the non-linear case. Indeed, although the idea is even the same, the governing equation is now

$$u_t + f(u)_x = 0. \tag{2.1.7}$$

This involves that the first time derivative of u is the first space derivative of $f(u)$, $u_t = -f(u)_x$; while the second time derivative is obtained by taking the time derivative of the governing equation (2.1.7). In practice,

$$u_{tt} = -(f(u)_x)_t = A(u)u_x, \tag{2.1.8}$$

where $A(u) = \nabla_u f$.

Substituting eq (2.1.8) in eq (2.1.1) we get

$$u(x_i, t+\Delta t) = u(x_i, t) - \Delta t f(u(x_i, t))_x + \frac{\Delta t^2}{2} A(u)u_x(x_i, t) + O(\Delta t^3) \tag{2.1.9}$$

If we discretize all the space derivatives with second order central formulas in the mesh point

x_i at time t_n we obtain

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{2\Delta x}\Big(f(u_{i+1}^n) - f(u_{i-1}^n)\Big) + \frac{\Delta t^2}{2\Delta x^2}\Bigg(A_{i+\frac{1}{2}}\Big(f(u_{i+1}^n) - f(u_i^n)\Big) - A_{i-\frac{1}{2}}\Big(f(u_i^n) - f(u_{i-1}^n)\Big)\Bigg), \tag{2.1.10}$$

where $A_{i\pm\frac{1}{2}}$ is the Jacobian matrix evaluated at $u_{i+\frac{1}{2}} = \frac{1}{2}(u_i^n + u_{i\pm1}^n)$, or the average of A between $A(u_i)$ and $A(u_{i\pm1})$.

The alternative conservative formulation of the Lax Wendroff scheme is:

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\Big(F_{i-\frac{1}{2}}^{\mathrm{LW}} - F_{i+\frac{1}{2}}^{\mathrm{LW}}\Big) \tag{2.1.11}$$

where

$$F_{i+\frac{1}{2}}^{\mathrm{LW}} = \frac{1}{2}\Big[f(u_{i+1}^n) + f(u_i^n) - \frac{\Delta t}{\Delta x}A_{i+\frac{1}{2}}\Big(f(u_{i+1}^n) - f(u_i^n)\Big)\Big].$$

Jacobian free schemes

In order to avoid the Jacobian matrix evaluation, first Richtmyer [102] (1967) and successively MacCormack [91] (1969) introduced some variant of the Lax-Wendroff scheme using a two steps method.

The first step, in the Richtmyer two-step Lax-Wendroff method, computes values for $f(u(x,t))$ at half time step $t_{n+\frac{1}{2}}$ and half grid point $x_{i+\frac{1}{2}}$. In the second step, the values at t_{n+1} depend on values computed at time $t_{n+\frac{1}{2}}$ and t_n. In practice, the first step or Lax step are:

$$u_{i+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\Big[u_{i+1}^n + u_i^n - \frac{\Delta t}{\Delta x}\Big(f(u_{i+1}^n) - f(u_i^n)\Big)\Big]; \tag{2.1.12}$$

$$u_{i-\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\Big[u_i^n + u_{i-1}^n - \frac{\Delta t}{\Delta x}\Big(f(u_i^n) - f(u_{i-1}^n)\Big)\Big]; \tag{2.1.13}$$

while, the second step is:

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\Big(f\Big(u_{i-\frac{1}{2}}^{n+\frac{1}{2}}\Big) - f\Big(u_{i+\frac{1}{2}}^{n+\frac{1}{2}}\Big)\Big). \tag{2.1.14}$$

Following the same idea of the two step, the MacCormack's method uses first forward differencing and then backward differencing in this way:

$$u_i^* = u_i^n - \frac{\Delta t}{\Delta x}\Big(f(u_{i+1}^n) - f(u_i^n)\Big); \tag{2.1.15}$$

$$u_i^{n+1} = \frac{1}{2}\Big[u_i^n + u_i^* - \frac{\Delta t}{\Delta x}\Big(f(u_i^*) - f(u_{i-1}^*)\Big)\Big]; \tag{2.1.16}$$

An alternative scheme is:

$$u_i^* = u_i^n - \frac{\Delta t}{\Delta x}\Big(f(u_i^n) - f(u_{i-1}^n)\Big); \tag{2.1.17}$$

$$u_i^{n+1} = \frac{1}{2}\Big[u_i^n + u_i^* - \frac{\Delta t}{\Delta x}\Big(f(u_{i+1}^*) - f(u_i^*)\Big)\Big]. \tag{2.1.18}$$

2.2 The High-Order Lax-Wendroff Method for Linear Problems

Let us consider the linear scalar equation (2.1.2)

$$u_t + a u_x = 0.$$

The high order Lax-Wendroff scheme is so set:

$$u_i^{n+1} = u_i^n + \sum_{k=1}^{m} \frac{(-1)^k c^k}{k!} \sum_{j=-p}^{p} \delta_{p,j}^k u_{i+j}^n, \tag{2.2.1}$$

where x_i are the nodes of a uniform mesh of step Δx; u_i^n is an approximation of the point value of the solution at x_i at the time $n\Delta t$, in which Δt represents the time step; $p \geq 1$ is a natural number; $c = \frac{a\Delta t}{\Delta x}$; and $\delta_{p,j}^k$ are the coefficients of the centered formula for the numerical approximation of the $k-$th derivative based on a $(2p+1)-$point stencil. In particular, these coefficients uniquely define the following formula:

$$f^{(k)}(x_i) \simeq D_p^k(f_i, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-p}^{p} \delta_{p,j}^k f(x_{i+j}), \tag{2.2.2}$$

such that

$$p_f^{(k)}(x_i) = \frac{1}{\Delta x^k} \sum_{j=-p}^{p} \delta_{p,j}^k f(x_{i+j}), \quad \forall f,$$

where p_f is the Lagrange interpolation polynomial characterized by

$$p_f^{(k)}(x_{i+j}) = f^{(k)}(x_{i+j}) \quad j = -p, \dots, p.$$

In this case, $f^{(k)}$ represents the $k-$th derivative of a single-variable f and imposing $f^{(0)} = f$. The numerical method (2.2.1) is obtained replacing the time derivatives by space derivative through the identities in numerical form

$$\partial_t^k u = (-1)^k a^k \partial_x^k u, \quad k = 1, 2, \dots$$

Remark 2.2.1 *The coefficients of formulas* (2.2.2) *do not depend on position i but just on Δx and order p. Of course, all coefficients need a uniform discretization.*

More properties and remarks concerning formulas and coefficients are treated on Appendix A-C.

2.3 Lax-Wendroff Approximate Taylor Scheme

The main difficulty to extend the Lax-Wendroff methods with high resolution to non-linear problems comes from the transformation of the time derivatives into the spatial derivatives. Many authors introduced different strategy to obtain high resolution in space and time. A successful technique for the spatial semi discretization was introduced with the ENO, WENO and CWENO approach [56, 67, 85, 86, 111, 112] which achieve arbitrarily high order spatial accuracy with excellent results in term of accuracy and performance.

A commonly used technique to obtain high order time discretization is the implementation of the multistage one-step *Strongly Stability Preserving Runge-Kutta schemes* [48]. A large class of numerical methods are introduced combining the ENO, WENO and CWENO approach for the spatial discretization with the SSP Runge-Kutta approach for the time discretization. Unfortunately, those combined methods are subjected to several restriction due to the stability properties not considering the difficulties in getting high order for the SSP Runge-Kutta schemes.

In order to simplify the formulation of high order accuracy in time, Qiu and Shu [99] developed a new time discretization following a Lax-Wendroff-type procedure and based

on the Cauchy-Kovaleskaya identity, where the numerical solution at a further time step is computed by a Taylor expansion in time with the time derivatives transformed into spatial derivatives trough the governing equation. The first important advantage of such scheme compared to the Runge-Kutta methods is that only one ENO or WENO reconstruction procedure flux splitting is necessary to be performed at each spatial position for each time step, regardless of the order of the method, yielding an overall better performance.

An alternative to avoid the CK procedure and the multistep or multistage one-step method, called Lax-Wendroff Approximate Taylor (LAT), has been proposed by Zorío, Mulet and Baeza in [132] based on an Approximate Taylor (AT) method.

2.3.1 LAT method

For the sake of simplicity, let us consider the one-dimensional system of conservation law

$$U_t + f(U)_x = 0, \tag{2.3.1}$$

with initial condition $U(x,0) = U_0(x)$; where $x \in \mathbb{R}^m$ and $U(x,t) : \mathbb{R}^m \times [0,+\infty) \to \mathbb{R}^d$ is a $d-$dimensional vector of conserved quantities.

Lax-Wendroff-type procedure

The generalized high order Lax-Wendroff method of order R for linear system of conservation law adopts the Taylor expansion in time to compute the numerical solution. In practice,

$$U_i^{n+1} = U_i^n + \sum_{k=1}^{R} \frac{(\Delta t)^k}{k!} U_i^{(k)} + \mathcal{O}\Big(\Delta t^{R+1}\Big), \tag{2.3.2}$$

where $\{x_i\}$ are the nodes of a uniform mesh of step Δx; $U_i^n \approx U(x_i, t_n)$ is a pointwise approximation of the solution at time $t_n = n\Delta t$ at position x_i; $U_i^{(k)}$ is an approximation of $\partial_t^k U(x_n, t_n)$.

Notation 2.3.1 *For this purpose, for every $k > 0$ let be:*

$$U_i^{(k)} = \partial_t^k U(x,t)\Big|_{x=x_i}^{t=t_n} + \mathcal{O}\Big(\Delta x^{R+1-k}\Big)$$
$$f_i^{(k)} = \partial_t^k f(U(x,t))\Big|_{x=x_i}^{t=t_n} + \mathcal{O}\Big(\Delta x^{R-k}\Big)$$

U solves the system of conservation law (2.3.1) then the time derivative can be written in term of spatial derivative. In particular, if the solutions U are assumed to be smooth enough

$$\partial_t^k U = -\partial_x \partial_t^{k-1} f(U). \tag{2.3.3}$$

Now, following the Faà di Bruno's formula [39] there exist some functions F_{k-1} depending on previous time derivatives of U, $F_{k-1} = F_{k-1}(U_i^n, U_i^{(1)}, \dots, U_i^{(l)})$ with $l < k$, such that $f_i^{(k-1)} = F_{k-1}$.

Theorem 2.3.1 (Faà di Bruno's formula) *Let $f : \mathbb{R}^m \to \mathbb{R}^p$ and $u : \mathbb{R} \to \mathbb{R}^m$ q times continuously differentiable functions. Then*

$$\frac{d^q f(u(t))}{dt^q} = \sum_{s \in \mathcal{P}_q} \binom{q}{s} f^{(|s|)}(u(t)) D^s u(t), \tag{2.3.4}$$

where $\mathcal{P}_q = \{s \in \mathbb{N}^q \text{ such that } \sum_{\mathrm{j}=1}^{\mathrm{j}} \mathrm{qs_j} = \mathrm{q}\}$, $|s| = \sum_{j=1}^q s_j$, and $D^s u(t)$ is an $m \times |s|$ matrix whose $(\sum_{l<j} s_l + i) - th$ column is given by

$$(D^s u(t))_{\sum_{l<j} s_l + i} = \frac{1}{j!} \partial_t^j u(x), \quad i = 1, \dots, s_j, \quad j = 1, \dots, q, \tag{2.3.5}$$

and the action of the k-th derivative tensor of f on a $m \times k$ matrix A is given by

$$f^{(k)}(u)A = \sum_{i_1, \dots, i_k}^{m} \frac{\partial^k f}{\partial u_{i_1}, \dots, \partial u_{i_k}}(u) A_{i_1,1} \dots A_{i_k,k} \in \mathbb{R}^p. \tag{2.3.6}$$

Thus, keeping in mind the Faà di Bruno's formula (2.3.4) and the Cauchy-Kovaleskaya identity (2.3.3) the Lax-Wendroff-type procedure is so set

1. Let $\{U_i^n\}_i$ be the pointwise data that approximate $U(x, t_n)$;

2. Let us compute the first time derivative of the solution U through the Cauchy-Kovalevskaya identity $U_t = -f(U)_x$. More specifically, looking for a scheme of order $2p$ we have

$$U_i^{(1)} = -[f(U)]_x \Big|_{x=x_i}^{t=t_n} = -\frac{\hat{f}_{i+\frac{1}{2}}^n - \hat{f}_{i-\frac{1}{2}}^n}{\Delta x} + \mathcal{O}(\Delta x^{2p-1}), \tag{2.3.7}$$

where $\hat{f}_{i+\frac{1}{2}}^n = \hat{f}(U_{i-p+1}^n, \dots, U_{i+p}^n)$ is computed with the upwind WENO numerical fluxes of order $2p-1$ in which $p = \lceil \frac{R+1}{2} \rceil$ [67];

3. Let us compute the second time derivative of the solution U using the Cauchy-Kovaleskaya identity and the Faà di Bruno's formula. In practice,

$$U_i^{(2)} = -[f(U)_t]_x\Big|_{x=x_i}^{t=t_n} = -\frac{f_{i+1}^{(1)} - f_{i-1}^{(1)}}{2\Delta x} + \mathcal{O}(\Delta x^{2p-2}), \tag{2.3.8}$$

where $f_{i+\frac{1}{2}}^{(1)} = F_1(U_i, U_i^{(1)}) = f'(U_i^n)U_i^{(1)}$;

... and so on until the selected order R.

At the end, once all the needed computation have been obtained, we can advance in time using the Taylor expansion in time and compute U_i^{n+1} as in Eq (2.3.2).

Remark 2.3.1 *A noteworthy observation is that the WENO method is needed only in Eq (2.3.7) to reconstruct the flux function on the staggered grid points.*

The Approximate Lax-Wendroff-type Procedure

Unfortunately, computationally speaking the Lax-Wendroff-type procedure is very expensive when the order R increases and it requires a large symbolic computation for each Faà di Bruno's formula.

Zorío, Baeza and Mulet proposed in [132] an alternative Lax-Wendroff-type procedure less expensive in computational sense and requiring only the knowledge of the flux function. Indeed, the rationale of their proposal lies that the exact computation of $\partial_t^{k-1} f(U)$ in the Lax-Wendroff-type procedure from Faà di Bruno's formula requires the knowledge of all the partial derivatives $\dfrac{\partial^l f}{\partial U_{i_1} \dots U_{i_l}}(U)$ for all $l < k$, but in general it is sufficient to know just the approximations of $\partial_t^{k-1} f(U)$. In fact, knowing $U_i^{(k)}$, trough the Taylor expansion and a family of central differences formula, we can compute $f_i^{(k)}$.

Notation 2.3.2 *In order to simplify the readability, let us introduce some notation:*

$$T_i^{k,n}(u) := \sum_{j=0}^{k} \frac{u_i^{(j)}}{j!}(t - t_n)^j;$$

$$\Delta_{\xi,i}^{p,q}(u) := \frac{1}{\Delta\xi^p} \sum_{j=-s}^{s} \beta_j^{p,q} u_j, \quad \text{where} \quad \mathrm{s} = \left\lfloor \frac{\mathrm{p}-1}{2} \right\rfloor + \mathrm{q}.$$

ξ is the direction in which we are working, for instance in the one-dimensional case $\Delta\xi \in \{\Delta x, \Delta t\}$. In other words, $T_i^{k,n}$ is the local Taylor expansion in time centered at $t = t_n$; while $\Delta_{\xi,i}^{p,q}$ is

the local centered finite differences operator that approximates the $p-th$ order derivatives to order $2q$ on grid with spacing $\Delta\xi$.

The Approximate Lax-Wendroff-type procedure proposed by Zorío, Baeza and Mulet [132] for a scheme of order R is so defined:

1. Let $\{U_i^n\}_i$ be the pointwise data that approximate $U(x_i, t_n)$;

2. Let us compute the first time derivative of the solution U trough the Cauchy-Kovalevskaya identity $U_t = -f(U)_x$. More specifically, looking for a scheme of order $2p$ we have

$$U_i^{(1)} = -[f(U)]_x\Big|_{x=x_i}^{t=t_n} = -\frac{\hat{f}^n_{i+\frac{1}{2}} - \hat{f}^n_{i-\frac{1}{2}}}{\Delta x} + \mathcal{O}(\Delta x^{2p-1}), \tag{2.3.9}$$

where $\hat{f}^n_{i+\frac{1}{2}} = \hat{f}(U^n_{i-p+1}, \ldots, U^n_{i+p})$ is computed with the upwind WENO numerical fluxes of order $2p-1$ in which $p = \lceil \frac{R+1}{2} \rceil$ [33, 67, 112];

$3+k-1$. For all $k = 1, \ldots, R-1$ let us define the $k-th$ order approximate Taylor polynomial $T_i^{k,n}$ as

$$T_i^{k,n}[U](t) = U_i^n + U_i^{(1)}(t-t_n) + \ldots + \frac{1}{k!}U_i^{(k)}(t-t_n)^k. \tag{2.3.10}$$

Hence, the $k-th$ order approximate time derivative of the flux is given by

$$f_i^{(k)} = \Delta_{t,i}^{k,\lceil \frac{R-k}{2} \rceil} f(T_i^{k,n}[U](t)), \tag{2.3.11}$$

$4+k-1$. One we know the local $k-th$ order approximate time derivative of the flux, let us compute the $(k+1)-th$ order time derivative of the solution U trough the Cauchy-Kovalevskaya identity $\partial_t^{k+1}U = -f^{(k)}(U)_x$. In practise, looking for a scheme of order $2p$ we have

$$U_i^{(k+1)} = -\Delta_{x,i}^{1,\lceil \frac{R-k}{2} \rceil}(f^{(k)}_{i+j}) + \mathcal{O}(\Delta x^{2p-1}), \tag{2.3.12}$$

where $j = -p, \ldots, p$ and $p = \lceil \frac{R-k}{2} \rceil$ ever for all $k = 1, \ldots, R-1$.

Thus , the numerical scheme of order R is

$$U_i^{n+1} = U_i^n + \sum_{k=1}^{R} \frac{(\Delta t)^k}{k!} U_i^{(k)}. \tag{2.3.13}$$

In order to simplify the aforementioned recursive procedure, we show in detail the numerical scheme of order 5. For this reason, let us consider a discretization of the space direction

using a uniform mesh grid of step Δx and the pointwise data that approximate $U(x_i, t_n)$, $\{U_i^n\}$; the first time derivative of the solution U of order 6 is given by:

$$U_i^{(1)} = -[f(U)]_x\Big|_{x=x_i}^{t=t_n} = -\frac{\hat{f}^n_{i+\frac{1}{2}} - \hat{f}^n_{i-\frac{1}{2}}}{\Delta x} + \mathcal{O}(\Delta x^5),$$

where $\hat{f}^n_{i+\frac{1}{2}} = \hat{f}(U^n_{i-2}, \ldots, U^n_{i+3})$ is being with the upwind WENO5 numerical fluxes [33, 67, 112].

Let us define the local first order Taylor approximation of U as

$$T_i^{1,n}[U](t) = U_i^n + U_i^{(1)}(t - t_n),$$

then, the local first order approximate time derivative of the flux is

$$\begin{aligned} f_i^{(1)} =& \Delta_{t,i}^{1,2} f(T_i^{1,n}[U](t)) = \\ =& \frac{f(T_i^{1,n}[U](t_{n+2})) - 8f(T_i^{1,n}[U](t_{n+1})) + 8f(T_i^{1,n}[U](t_{n-1})) - f(T_i^{1,n}[U](t_{n-2}))}{12\Delta t}. \end{aligned}$$

The local second time derivative of the solution U of order 4 is so defined:

$$U_i^{(2)} = -\Delta_{x,i}^{1,2}(f_{i+j}^{(1)}) + \mathcal{O}(\Delta x^3) = -\frac{f_{i-2}^{(1)} - 8f_{i-1}^{(1)} + 8f_{i+1}^{(1)} - f_{i+2}^{(1)}}{12\Delta x}.$$

The local second order Taylor approximation of U is:

$$T_i^{2,n}[U](t) = U_i^n + U_i^{(1)}(t - t_n) + \frac{1}{2}U_i^{(2)}(t - t_n)^2,$$

then, the local second order approximate time derivative of the flux is given by:

$$\begin{aligned} &f_i^{(2)} = \Delta_{t,i}^{2,2} f(T_i^{2,n}[U](t)) = \\ &= \frac{-f(T_i^{2,n}[U](t_{n+2})) + 16f(T_i^{2,n}[U](t_{n+1})) - 3f(T_i^{2,n}[U](t_n)) + 16f(T_i^{2,n}[U](t_{n-1})) - f(T_i^{2,n}[U](t_{n-2}))}{12\Delta t^2}. \end{aligned}$$

The local third time derivative of the solution U of order 4 is so defined:

$$U_i^{(3)} = -\Delta_{x,i}^{1,2}(f_{i+j}^{(2)}) + \mathcal{O}(\Delta x^3) = -\frac{f_{i-2}^{(2)} - 8f_{i-1}^{(2)} + 8f_{i+1}^{(2)} - f_{i+2}^{(2)}}{12\Delta x}.$$

The third order Taylor approximation of U is:

$$T_i^{3,n}[U](t) = U_i^n + U_i^{(1)}(t-t_n) + \frac{1}{2}U_i^{(2)}(t-t_n)^2 + \frac{1}{6}U_i^{(3)}(t-t_n)^3,$$

consequently, the local third order approximate time derivative of the flux is given by:

$$\begin{aligned} f_i^{(3)} =& \Delta_{t,i}^{3,1} f(T_i^{3,n}[U](t)) = \\ =& \frac{f(T_i^{3,n}[U](t_{n+2})) - 8f(T_i^{3,n}[U](t_{n+1})) + 8f(T_i^{3,n}[U](t_{n-1})) - f(T_i^{3,n}[U](t_{n-2}))}{2\Delta t^3}. \end{aligned}$$

The fourth time derivative of the solution U of order 2 is so set:

$$U_i^{(4)} = -\Delta_{x,i}^{1,1}(f_{i+j}^{(3)}) + \mathcal{O}(\Delta x^1) = -\frac{f_{i+1}^{(3)} - f_{i-1}^{(3)}}{2\Delta x}.$$

The local fourth order Taylor approximation of U is so get:

$$T_i^{4,n}[U](t) = U_i^n + U_i^{(1)}(t-t_n) + \frac{1}{2}U_i^{(2)}(t-t_n)^2 + \frac{1}{6}U_i^{(3)}(t-t_n)^3 + \frac{1}{24}U_i^{(4)}(t-t_n)^4,$$

thus, the fourth order approximate time derivative of the flux is defined as:

$$\begin{aligned} f_i^{(4)} &= \Delta_{t,i}^{4,1} f(T_i^{4,n}[U](t)) = \\ &= \frac{f(T_i^{4,n}[U](t_{n+2})) - 4f(T_i^{4,n}[U](t_{n+1})) + 6f(T_i^{4,n}[U](t_n)) - 4f(T_i^{4,n}[U](t_{n-1})) + f(T_i^{4,n}[U](t_{n-2}))}{\Delta t^4}. \end{aligned}$$

The last local time derivative of U is given by:

$$U_i^{(5)} = -\Delta_{x,i}^{1,1}(f_{i+j}^{(4)}) + \mathcal{O}(\Delta x^1) = -\frac{f_{i+1}^{(4)} - f_{i-1}^{(4)}}{2\Delta x}.$$

Finally, the LAT scheme of order 5 written in non conservative form is as follow:

$$U_i^{n+1} = U_i^n + \Delta t U_i^{(1)} + \frac{\Delta t^2}{2}U_i^{(2)} + \frac{\Delta t^3}{6}U_i^{(3)} + \frac{\Delta t^4}{24}U_i^{(4)} + \frac{\Delta t^5}{120}U_i^{(5)}.$$

Zorío et al [132] proved that the LAT scheme defined by (2.3.13) is $R-th$ order accurate. Furthermore, in [132] is also proven that the scheme (2.3.13) could be written in conservative form.

Remark 2.3.2 *The LAT method* (2.3.13) *is not a properly generalization of the high order*

Lax-Wendroff scheme (2.3.2) *in the sense that it is not reduced to linear Lax-Wendroff method when* $f(U) = aU$. *In fact, let us consider* $R = 2$ *and a linear case* $f(U) = aU$; *the first time derivative of* U *is:*

$$U_i^{(1)} = -\frac{U_{i+1}^n - U_{i-1}^n}{2\Delta x};$$

while

$$U_i^{(2)} = -\frac{U_{i+2}^n - 2U_i^n + U_{i-2}^n}{4\Delta x^2}.$$

Then, the LAT scheme (2.3.13) *of order 2 applied to the linear case return:*

$$U_i^{n+1} = U_i^n - \frac{a\Delta t}{2\Delta x}\Big(U_{i+1}^n - U_{i-1}^n\Big) - \frac{a^2\Delta t^2}{8\Delta x^2}\Big(U_{i+2}^n - 2U_i^n + U_{i-2}^n\Big).$$

This method is different from the Lax-Wendroff scheme (2.1.5) *and it adopt* $(4R+1)-$*points losing the stability property of the standard Lax-Wendroff method. (see [84]), therefore it uses a wider stencil of Lax-Wendroff type schemes of the same order.*

As already mentioned, it is desirable to construct Taylor based methods with optimal stencil, which reduce to the Lax-Wendroff schemes when applied to linear systems.

Chapter 3

Adaptive Compact Approximate Taylor Method for systems of conservation law

As we have seen, the extension of the Lax-Wendroff methods to non-linear systems of conservation law is not immediately. Many authors have developed numerical methods that use Lax-Wendroff-type approach for the time discretization as an alternative to multistep or multistage one-step schemes like the SSP Runge-Kutta methods (see [48]): this is the case of the original finite volume ENO schemes (see [56]); or the approach followed by Toro and collaborators in the design of the so-called ADER (arbitrary high order schemes using higher order derivatives) methods (see [108, 118, 121]); or the approach proposed by Qiu and Shu in [99]. A numerical alternative to those methods has been proposed in [132] based on an Approximate Taylor (AT) method. In this case, the time derivative are approximated using the high-order centered differentiation formulas combined with Taylor approximations in time that are computed in a recursive way. Nevertheless, AT schemes are not exactly a generalization of Lax-Wendroff methods, indeed they have $(4p+1)-$point stencils and worse linear stability properties than the original Lax-Wendroff schemes. Despite that, they can be stabilized by using one WENO reconstruction for spatial cell and time step, as in [99] and the resulting methods give good results usually under CFL$= 0.5$ condition. The focus of this chapter is present an adaptive family of numerical methods, named ACAT schemes, for non-linear systems of conservation law based on an approximate Taylor procedure that constitute a proper generalization of Lax-Wendroff methods, i.e. that reduce to the standard high-order

Lax-Wendroff methods when the flux is linear. Thus building a family of high-order numerical methods L^2-stable under CFL-1 condition. As it expected for Lax-Wendroff schemes, such a family of methods would lead to spurious oscillations near discontinuity [55] and a numerical technique would be required to avoid them as: flux limiters [70, 116]; essentially non-oscillatory reconstructions like ENO [56] or WENO [111, 112] or CWENO [85, 86]; adaptive approach [14]; MOOD approach [26]; other shock capturing techniques. In particular, we will focus on an order adaptive version that are able to avoid the spurious oscillations according with a family of numerical high order smoothness indicators.

3.1 Compact Approximate Taylor Method

Carrillo and Parés in [15] designed a compact variant of the LAT scheme (2.3.13) that properly generalize the Lax-Wendroff methods for linear systems (2.3.2). These methods are based on the conservative expression of the LAT scheme with the difference that the numerical flux $F^p_{i+\frac{1}{2}}$ is computing using only the values

$$U^n_{i-p+1}, \dots, U^n_{i+p},$$

where $2p$ is the order of accuracy. In this way, the numerical solution U^n_i is updated using only the values at the centered $(2p+1)-$point stencil. For this reason, let us consider the one-dimensional system of conservation law

$$U_t + f(U)_x = 0,$$

with initial condition $U(x,0) = U_0(x)$, where $U : \mathbb{R} \times [0, +\infty) \to \mathbb{R}^d$ is a $d-$dimensional vector of conserved quantities, and $f : \mathbb{R}^d \to \mathbb{R}^d$ is the flux function.
As we have seen on previous section, the generalized Lax-Wendroff method is used to update the numerical solution:

$$U^{n+1}_i = U^n_i + \sum_{k=1}^{2P} \frac{(\Delta t)^k}{k!} U^{(k)}_i,$$

where $\{x_i\}$ are the nodes of a uniform mesh of step Δx; U^n_i is an approximation of the value of the exact solution $U(x,t)$ at time $t_n = n\Delta t$ at position x_i [58]; and $U^{(k)}_i$ is an approximation of $\partial^k_t U(x_i, t_n)$, where the $k-th$derivative in time of U are computed with a compact numerical version of the Cauchy-Kovalesky procedure introduced by Carrillo and Parés in

[15].

The final expression of the $2P-$order Compact Approximate Taylor (CAT) method in conservative form is:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F^P_{i-\frac{1}{2}} - F^P_{i+\frac{1}{2}}\Big), \tag{3.1.1}$$

where the flux functions $F^P_{i\pm\frac{1}{2}}$ are computed, respectively, on stencil $S^P_{i\pm\frac{1}{2}}$; in which

$$S^P_{i+\frac{1}{2}} = \{U^n_{i-P+1}, \dots, U^n_{i+P}\};$$

$$F_{i+\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} f^{(k-1)}_{i+\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,\frac{1}{2}}\Big(f^{(k-1)}_{i,*}, \Delta x\Big) \tag{3.1.2}$$

and

$$f^{(k-1)}_{i+\frac{1}{2}} = A_P^{0,\frac{1}{2}}\Big(f^{(k-1)}_{i,*}, \Delta x\Big) = \sum_{j=-P+1}^{P} \gamma^{0,\frac{1}{2}}_{P,j} f^{(k-1)}_{i+j}$$

is an interpolatory formulas of order $2P-1$ based on $2P-$point stencil.

In order to simplify the readability of the scheme, let focus on the second order ($P = 1$) CAT2 method.

3.1.1 CAT2

In this case,

- The relative stencils are:

$$S^1_{i+\frac{1}{2}} = \{U^n_i, U^n_{i+1}\} \quad \text{and} \quad S^1_{i-\frac{1}{2}} = \{U^n_{i-1}, U^n_i\};$$

- The flux reconstructions are:

$$F^1_{i+\frac{1}{2}} = f^{(0)}_{i+\frac{1}{2}} + \frac{\Delta t}{2} f^{(1)}_{i+\frac{1}{2}}, \tag{3.1.3}$$

$$F^1_{i-\frac{1}{2}} = f^{(0)}_{i-\frac{1}{2}} + \frac{\Delta t}{2} f^{(1)}_{i-\frac{1}{2}}; \tag{3.1.4}$$

  where

  1. $f^{(0)}_{i+\frac{1}{2}} = \frac{1}{2}\Big(f^n_i + f^n_{i+1}\Big)$ and $f^{(0)}_{i-\frac{1}{2}} = \frac{1}{2}\Big(f^n_i + f^n_{i-1}\Big)$ are the interpolations of the stage values at time t_n;

2. $f^{(1)}_{i+\frac{1}{2}} = \frac{1}{2}\left(f^{(1)}_i + f^{(1)}_{i+1}\right)$ and $f^{(1)}_{i-\frac{1}{2}} = \frac{1}{2}\left(f^{(1)}_i + f^{(1)}_{i-1}\right)$ are the interpolations of the first time derivative of the flux at time t_n;

   a1. $f^{(1)}_{i-1+j} = \frac{1}{\Delta t}\left(f\left(U^n_{i-1+j} + \Delta t U^{(1)}_{i-1,j}\right) - f^n_{i-1+j}\right)$ for $j = 0, 1$ are the first time derivatives of flux computed in point stencil $S^1_{i-\frac{1}{2}}$;

   a2. $f^{(1)}_{i+j} = \frac{1}{\Delta t}\left(f\left(U^n_{i+j} + \Delta t U^{(1)}_{i,j}\right) - f^n_{i+j}\right)$ for $j = 0, 1$ are the first time derivatives of flux computed in point stencil $S^1_{i+\frac{1}{2}}$;

   a3. $U^{(1)}_{i-1,j} = -\frac{1}{\Delta x}\left(f^n_i - f^n_{i-1}\right)$ for $j = 0, 1$ are the first time derivatives of the solution U at time t_n for each position of stencil $S^1_{i-\frac{1}{2}}$ necessary to compute the Taylor expansion truncated at first term;

   a4. $U^{(1)}_{i,j} = -\frac{1}{\Delta x}\left(f^n_{i+1} - f^n_i\right)$ for $j = 0, 1$ are the first time derivatives of the solution U at time t_n for each position of stencil $S^1_{i+\frac{1}{2}}$ necessary to compute the Taylor expansion truncated at first term.

Finally, we find that the flux reconstructions are so defined:

$$F^1_{i+\frac{1}{2}} = \frac{1}{4}\left(f^n_i + f^n_{i+1} + f\left(U^n_i + \Delta t U^{(1)}_{i,0}\right) + f\left(U^n_{i+1} + \Delta t U^{(1)}_{i,1}\right)\right); \tag{3.1.5}$$

$$F^1_{i-\frac{1}{2}} = \frac{1}{4}\left(f^n_i + f^n_{i-1} + f\left(U^n_i + \Delta t U^{(1)}_{i-1,1}\right) + f\left(U^n_{i-1} + \Delta t U^{(1)}_{i-1,0}\right)\right). \tag{3.1.6}$$

So the idea behind the algorithm is:

Step 1: Compute $f^{(0)}_{i+\frac{1}{2}}$ adopting an interpolatory formula on stencil $S^1_{i+\frac{1}{2}}$;

Step 2: Compute the first derivatives in time through the numerical compact Cauchy-Kovalesky $\partial_t U = -\partial_x f^n$ as done in step a3. and a4.;

Step 3: Compute the Taylor expansions truncated at first term $U^{1,n+1}_{i,j} = U^n_{i+j} + \Delta t U^{(1)}_{i,j}$;

Step 4: Compute the first time derivatives of flux using the first difference formulas $f^{(1)}_{i,j} = \frac{1}{\Delta t}(f(U^{1,n+1}_{i,j}) - f^n_{i+j})$;

Step 5: Compute $f^{(1)}_{i+\frac{1}{2}}$ through $f^{(1)}_{i,j}$ adopting an interpolatory formula on stencil $S^1_{i+\frac{1}{2}}$;

Step 6: Compute $F_{i+\frac{1}{2}}$ as Taylor expansion $F_{i+\frac{1}{2}} = f^{(0)}_{i+\frac{1}{2}} + \frac{\Delta t}{2} f^{(1)}_{i+\frac{1}{2}}$.

Figure 3.1.1 could help to have a graphic idea of the necessary stencil to compute the right flux reconstruction $F^1_{i+\frac{1}{2}}$.

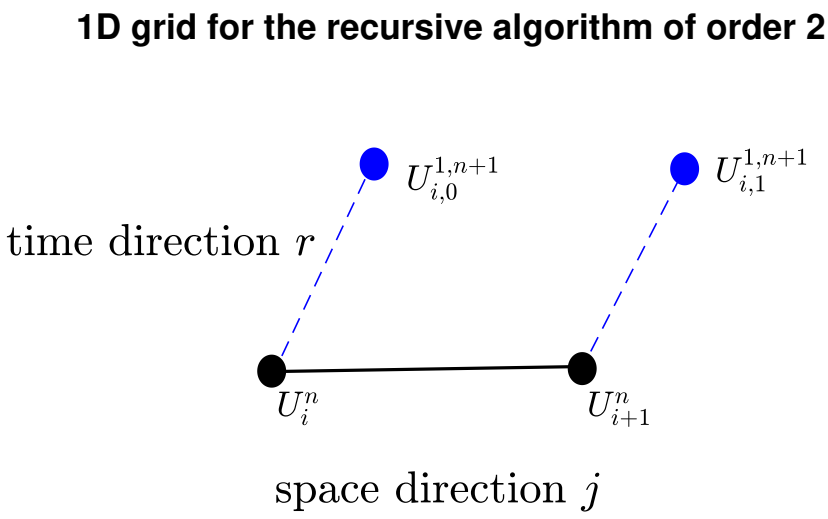

Figure 3.1.1: Local ghost grid for right flux reconstruction on CAT2

3.1.2 Numerical comparison between CAT2 and the Lax-Wendroff-Richtmyer-McCormack schemes

In this section we focus on the behaviour of the CAT2 procedure applied to several 1D problems: the 1D linear transport equation and Burgers equation,compared with the two-step Lax-Wendroff extension: Ritchmyer and McCormack.

1D scalar transport equation

Let us consider the linear scalar transport equation

$$u_t + u_x = 0. \tag{3.1.7}$$

We solve it with different types of initial conditions including smooth and no-smooth condition.

Test 1: We consider the transport equation (3.1.7) with smooth initial condition:

$$u_0(x) = \frac{1}{2}\sin(\pi x). \tag{3.1.8}$$

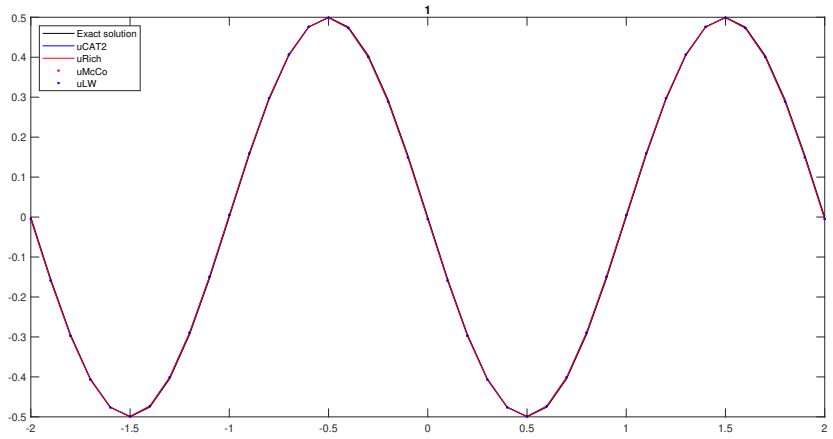

Figure 3.1.2: Test 1: Transport equation with initial condition (3.1.8). Exact and numerical solutions at $t = 1$ obtained with CAT2, Lax-Wendroff, Ritchmyer and McCormack schemes.

	CAT2		Lax-Wendroff		Ritchmyer		McCormack	
Points	Error	Order	Error	Order	Error	Order	Error	Order
20	1.07E-2	-	1.07E-2	-	1.07E-2	-	1.07E-2	-
40	2.61E-3	2.03	2.61E-3	2.03	2.61E-3	2.03	2.61E-3	2.03
80	6.41E-4	2.02	6.41E-4	2.02	6.41E-4	2.02	6.41E-4	2.02
160	1.56E-4	2.04	1.56E-4	2.04	1.56E-4	2.04	1.56E-4	2.04

Table 3.1: Test 1: Transport equation with smooth initial condition (3.1.8). Errors and numerical rates at $t = 2$ obtained with CAT2, Lax-Wendroff, Ritchmyer and McCormack schemes.

Test 2: We consider the transport equation (3.1.7) with non-smooth initial condition:

$$u_0(x) = \begin{cases} 2 & \text{if } \frac{1}{5} < x <= \frac{7}{5}; \\ 1 & \text{if } 0 < x <= \frac{1}{5} \; \frac{7}{5} < x <= 2; \\ 0 & \text{otherwise.} \end{cases} \quad . \tag{3.1.9}$$

We solve numerically the scalar equation with smooth (3.1.8) and non-smooth (3.1.9) initial condition in the interval $[-2, 2]$, using 40 mesh point, CFL= 0.9, and periodic boundary conditions. As we can see, Figures 3.1.2-3.1.3 and Table 3.1, prove that, even if we have smooth or non-smooth initial conditions, all the methods are exactly the same and they are an exact extension of the Lax-Wendroff scheme.

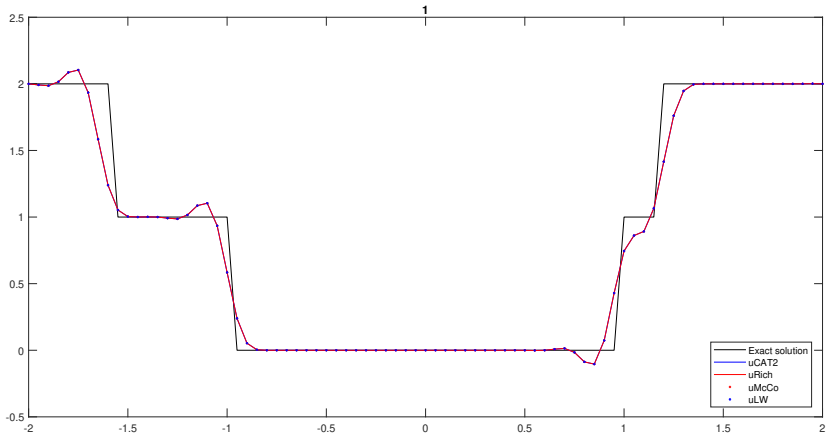

Figure 3.1.3: Test 2: Transport equation with initial condition (3.1.9). Exact and numerical solutions at $t = 1$ obtained with CAT2, Lax-Wendroff, Ritchmyer and McCormack schemes.

1D scalar Burgers equation

Let us consider the burgers scalar equation

$$u_t + \frac{1}{2}(u^2)_x = 0. \tag{3.1.10}$$

We solve it with different types of initial conditions including smooth and no-smooth conditions.

Test 3: We consider the burgers equation (3.1.10) with smooth initial condition:

$$u_0(x) = \frac{1}{2}\sin(\pi x). \tag{3.1.11}$$

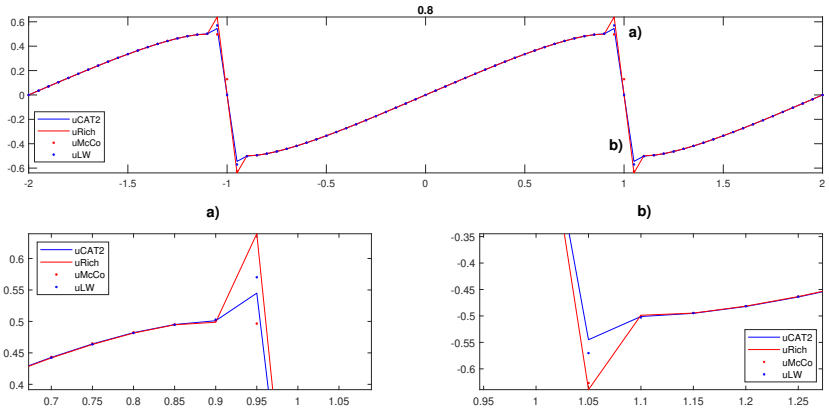

Figure 3.1.4: Test 3: Burgers equation with initial condition (3.1.11). Numerical solutions at $t = 0.8$ obtained with CAT2, Lax-Wendroff, Ritchmyer and McCormack schemes.

Test 4: We consider Burgers equation (3.1.10) with non-smooth initial condition:

$$u_0(x) = \begin{cases} 2 & \text{if } \frac{1}{5} < x <= \frac{7}{5}; \\ 1 & \text{if } 0 < x <= \frac{1}{5}\ \frac{7}{5} < x <= 2; \\ 0 & \text{otherwise.} \end{cases} \quad . \tag{3.1.12}$$

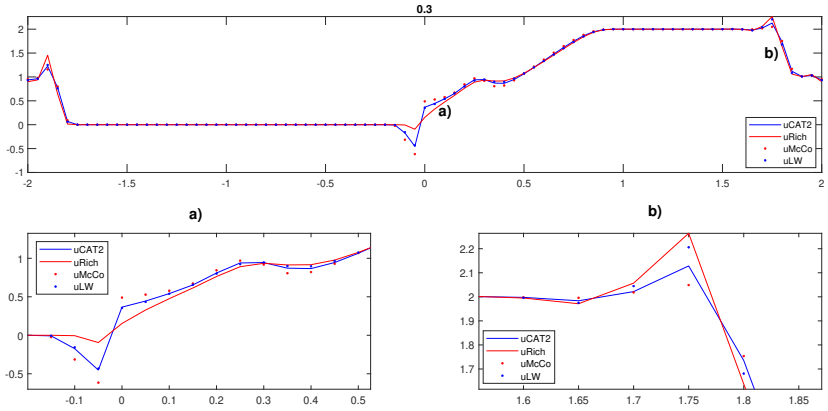

Figure 3.1.5: Test 4: Transport equation with initial condition (3.1.12). Numerical solutions at $t = 0.3$ obtained with CAT2, Lax-Wendroff, Ritchmyer and McCormack schemes.

We solve numerically the burgers equation with smooth (3.1.11) and non-smooth (3.1.12) initial conditions in the interval $[-2, 2]$, using 80 mesh point, CFL$= 0.9$, and periodic boundary conditions. As we can see, Figures 3.1.4-3.1.5, prove that, even if we have smooth or non-smooth initial condition, the CAT2 reconstruction introduce less spourious oscillations close to the discontinuities and in some part it is very similar to the Lax-Wendroff scheme.

3.1.3 CAT2P

As we have seen, equation (3.1.1) and (3.1.2) describe the $2P-$order Compact Approximate Taylor method in conservative form. This method was designed to be a properly extension of the $2P-$order Lax-Wendroff method for linear system what implies the linear stability for these methods under the usual CFL-1 condition, see Appendix B.

Since the numerical differentiation formulas play a major role in the algorithm, let us introduce some notation to describe the formulas that will be used. In the same way we have quickly seen above, the operator $A_P^{k,q}$ represents the interpolatory formula of order $2P-k$ that approximates the $k-$th derivative of a function at the point $x_i + q\Delta x$ using its values at the $2P-$point stencil $S = \{x_{\lfloor i+q \rfloor - P+1}, \ldots, x_{\lfloor i+q \rfloor + P}\}$. For this reason, in order to make

more readability the notation we will suppose $|q| < 1$, $A_P^{k,q}$ is so defined

$$f^{(k)}(x_i + q\Delta x) \approx A_P^{k,q}(f, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-P+1}^{P} \gamma_{P,j}^{k,q} f(x_{i+j}). \tag{3.1.13}$$

In case $k = 0$, we have

$$f^{(0)}(x_i + q\Delta x) \approx A_P^{0,q}(f, \Delta x) = \sum_{j=-P+1}^{P} \gamma_{P,j}^{0,q} f(x_{i+j})$$

that means the values at $x_i + q\Delta x$ of the Lagrange polynomial that interpolates the values of f at time t_n at the points $x_{i-P+1}, \ldots, x_{i+P}$.

Remark 3.1.1 *The coefficients $\gamma_{P,j}^{k,q}$ of the differential formulas can be computed by a recursive procedure introduced in [15, 41, 42]. See also Appendix A-C for more details.*

Continuing with notation, the following one

$$A_P^{k,q}(f_*, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-P+1}^{P} \gamma_{P,j}^{k,q} f_{i+j} \tag{3.1.14}$$

will be used to indicate that the formula is applied to some approximations f_i of f and not to its exact values $f(x_i)$. In case where there are two or more indices, the symbol $*$ will be used to indicate with respect to which the differentiation formula is applied. Indeed, the next approximations will be used, from now on, to computed the numerical fluxes:

$$\partial_t^k U(x_{i+j}, t_n) \approx -A_P^{1,j}\left(f_{i,*}^{(k)}, \Delta x\right) = -\frac{1}{\Delta x} \sum_{\ell=-P+1}^{P} \gamma_{P,\ell}^{1,j} f_{i,\ell}^{(k)}, \tag{3.1.15}$$

$$\partial_t^k f(U)(x_{i+j}, t_n) \approx A_P^{k,0}\left(f_{i,j}^{k,*}, \Delta t\right) = \frac{1}{\Delta t^k} \sum_{r=-P+1}^{P} \gamma_{P,r}^{k,0} f_{i,j}^{k,n+r}, \tag{3.1.16}$$

$$\partial_t^k f(U)\left(x_i + \frac{\Delta x}{2}, t_n\right) \approx A_P^{0,\frac{1}{2}}\left(f_{i,*}^{(k)}, \Delta x\right) = \sum_{j=-P+1}^{P} \gamma_{P,j}^{0,\frac{1}{2}} f_{i,j}^{(k)}. \tag{3.1.17}$$

In (3.1.15), numerical differentiation in space is used to approximate the time derivative of the solution at x_{i+j} from the local approximations $f_{i,\ell}^{(k)}$, $\ell = -P+1, \ldots, P$ according to (2.3.3). In (3.1.16), numerical differentiation in time is used to approximate the $k-$th time derivative of $f(U)$ in position x_{i+j} at time t_n from some approximations $f_{i,j}^{k,n+r}$ of $f(U)(x_{i+j}, t_{n+r})$, $r = -P+1, \ldots, P$ computed with the, respectively, Taylor expansion truncated at term k.

Finally, in (3.1.17), Lagrange interpolation is used to approximate the value of the $k-$th time derivative of $f(U)$ at position $x_i + \Delta x/2$ at time t_n from $f_{i,\ell}^{(k)}$, $\ell = -P+1, \ldots, P$.

Adopting the above notation, the final expression of the right numerical flux of order $2P$ is so get:

$$F_{i+\frac{1}{2}}^{P} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,\frac{1}{2}} \left(f_{i,*}^{(k-1)}, \Delta x \right) = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} \sum_{j=-P+1}^{P} \gamma_{P,j}^{0,\frac{1}{2}} f_{i,j}^{(k-1)}, \tag{3.1.18}$$

where the high order time derivative of the flux are computed following and extending the iterative algorithm presented for CAT2 in above section (see [13–15] for more details):

1. Define $f_{i,j}^{(0)} := f(U_{i+j}^n)$ for all $j = -P+1, \ldots, P$;
2. For every $k = 1, \ldots, 2P-1$ act in this way:
   (a) Compute the $k-$th derivative of U at time step t_n for each position x_{i+j} with $j = -P+1, \ldots, P$ through the numerical compact version of the Cauchy-Kovalesky identity (2.3.3) as:
   $$U_{i,j}^{(k)} = -A_P^{1,j} \left(f_{i,*}^{(k-1)}, \Delta x \right);$$
   (b) Compute the Taylor expansion of U in time truncated at term k for all positions x_{i+j} with $j = -P+1, \ldots, P$ at time t_{n+r} with $r = -P+1, \ldots, P$ as:
   $$U(x_{i+j}, t_{n+r}) \approx U_{i,j}^{k,n+r} = U_{i+j}^n + \sum_{m=1}^{k} \frac{(r\Delta t)^m}{m!} U_{i,j}^{(m)};$$
   (c) Compute the $k-$th time derivative of flux for each position x_{i+j} with $j = -P+1, \ldots, P$ at time t_n as:
   $$f_{i,j}^{(k)} = A_P^{k,j} \left(f_{i,j}^{k,*}, \Delta t \right),$$
   where $f_{i,j}^{k,*}$ means that we are applying the A operator in time and in particular we apply the differentiation formula to
   $$f_{i,j}^{k,n-P+1}, \ldots, f_{i,j}^{k,n+P}$$
   in which $f_{i,j}^{k,n+r}$ represents $f\left(U_{i,j}^{k,n+r}\right)$ for all $j, r = -P+1, \ldots, P$.

Remark 3.1.2 *Observe that the computation of the numerical flux $F_{i+\frac{1}{2}}^{P}$ requires the approximation of U at the nodes of a space-time grid of $2P \times 2P$ points: $U_{i,j}^{k,n+r}$, $j, r = -P+1, \ldots, P$*

(see Figure 3.1.6). The approximations of the solution u at times $(n-P+1)\Delta t, \ldots, (n-1)\Delta t$ *are different from the ones already computed in the previous steps:* $U_{i+j}^{n-P}, \ldots, U_{i+j}^{n-1}$. *In other words, the discretization in time is not based on a multistep method but in a one-step one: in fact it can be interpreted as a RK method whose stages are* $\tilde{U}_{i,j}^{n+r}$, $r = -P, \ldots, P$: *see [13–15].*

Remark 3.1.3 *These approximations are local in sense: let us suppose that* $i_1+j_1 = i_2+j_2 = \ell$, *i.e.* x_ℓ *belongs to* $S^P_{i_1+\frac{1}{2}}$ *and* $S^P_{i_2+\frac{1}{2}}$ *with local coordinates* j_1 *and* j_2 *respectively. Then,* $f^{(k)}_{i_1,j_1}$ *and* $f^{(k)}_{i_2,j_2}$ *are, in general, two different approximations of* $\partial_t^k f(U)(x_\ell, t_n)$.

Remark 3.1.4 *All the properties concerning the CAT2P, i.e. stability, accuracy and consistence, will be show in Appendix A and B.*

CAT4

In order to help the readability of the general iterative algorithm used to compute the $k-$th time derivative of flux, we will be shown also the practical way to apply CAT4 method. Unlike the CAT2 method CAT4 scheme will give a clearer view of the high-order CAT automatism by introducing in detail the development of the iterative procedure behind the computing of the time derivatives of the flux. With this in mind, see Figure 3.1.6 to have a graphic idea of the necessary stencil. The iterative CAT4 algorithm is so defined:

Step 1: Compute $f^{(0)}_{i+\frac{1}{2}}$ adopting the interpolatory formula on stencil $S^2_{i+\frac{1}{2}}$ as:

$$f^{(0)}_{i+\frac{1}{2}} = \sum_{j=-1}^{2} \gamma^{0,\frac{1}{2}}_{2,j} f^n_{i+j};$$

Step 2: Compute the first time derivative of U at time t_n at position x_{i+j} with $j = -1, \ldots, 2$ through the numerical compact Cauchy-Kovalesky identity (2.3.3) as:

$$U^{(1)}_{i,j} = -\frac{1}{\Delta x} \sum_{s=-1}^{2} \gamma^{1,j}_{2,s} f^n_{i+s};$$

Step 3: Compute the Taylor expansion in time truncated at first term at time t_{n+r} with $r =$

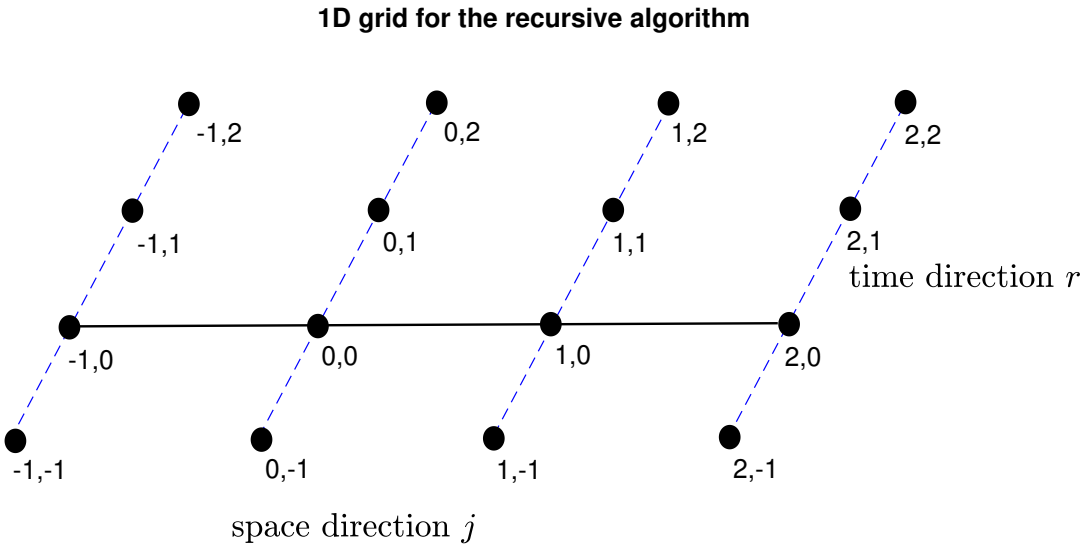

Figure 3.1.6: Local space-time grid where approximations of U are computed to calculate $F^P_{i+1/2}$ with $P = 2$. For simplicity a pair j, r represents the point (x_{i+j}, t_{n+r}). Taylor expansions in time are used to obtain these approximations following the blue lines. These Taylor expansions are centered in the points lying on the black line.

$-1, \ldots, 2$ for each position x_{i+j} with $j = -1, \ldots, 2$ as:

$$U^{1,n+r}_{i,j} = U^n_{i+j} + r\Delta t U^{(1)}_{i,j};$$

Step 4: Compute the first time derivative of f at time t_n at position x_{i+j} with $j = -1, \ldots, 2$ using the four fluxes $f^{1,n+r}_{i,j}$ as:

$$f^{(1)}_{i,j} = \frac{1}{\Delta t} \sum_{r=-1}^{2} \gamma^{1,0}_{2,r} f^{1,n+r}_{i,j};$$

Step 5: Compute $f^{(1)}_{i+\frac{1}{2}}$ adopting the interpolatory formula on stencil $S^2_{i+\frac{1}{2}}$ as:

$$f^{(1)}_{i+\frac{1}{2}} = \sum_{j=-1}^{2} \gamma^{0,\frac{1}{2}}_{2,j} f^{(1)}_{i,j};$$

Step 6: Compute the second time derivative of U at time t_n and position x_{i+j} with $j = -1, \ldots, 2$ from the first time derivatives of f using the (2.3.3) as:

$$U^{(2)}_{i,j} = -\frac{1}{\Delta x} \sum_{s=-1}^{2} \gamma^{1,j}_{2,s} f^{(1)}_{i,s};$$

Step 7: Compute the Taylor expansion in time truncated at second term at time t_{n+r} with $r = -1, \ldots, 2$ for each position x_{i+j} with $j = -1, \ldots, 2$ as:

$$U_{i,j}^{2,n+r} = U_{i+j}^{n} + r\Delta t U_{i,j}^{(1)} + \frac{(r\Delta t)^2}{2} U_{i,j}^{(2)};$$

Step 8: Compute the second time derivative of f at time t_n at position x_{i+j} with $j = -1, \ldots, 2$ using the four fluxes $f_{i,j}^{2,n+r}$ as:

$$f_{i,j}^{(2)} = \frac{1}{\Delta t^2} \sum_{r=-1}^{2} \gamma_{2,r}^{2,0} f_{i,j}^{2,n+r};$$

Step 9: Compute $f_{i+\frac{1}{2}}^{(2)}$ adopting the interpolatory formula on stencil $S_{i+\frac{1}{2}}^{2}$ as:

$$f_{i+\frac{1}{2}}^{(2)} = \sum_{j=-1}^{2} \gamma_{2,j}^{0,\frac{1}{2}} f_{i,j}^{(2)};$$

Step 10: Compute the third time derivative of U at time t_n and position x_{i+j} with $j = -1, \ldots, 2$ from the second time derivatives of f using the (2.3.3) as:

$$U_{i,j}^{(3)} = -\frac{1}{\Delta x} \sum_{s=-1}^{2} \gamma_{2,s}^{1,j} f_{i,s}^{(2)};$$

Step 11: Compute the Taylor expansion in time truncated at third term at time t_{n+r} with $r = -1, \ldots, 2$ for each position x_{i+j} with $j = -1, \ldots, 2$ as:

$$U_{i,j}^{3,n+r} = U_{i+j}^{n} + r\Delta t U_{i,j}^{(1)} + \frac{(r\Delta t)^2}{2} U_{i,j}^{(2)} + \frac{(r\Delta t)^3}{6} U_{i,j}^{(3)};$$

Step 12: Compute the third (it is also the last derivative that we can compute) time derivative of f at time t_n at position x_{i+j} with $j = -1, \ldots, 2$ using the four fluxes $f_{i,j}^{3,n+r}$ as:

$$f_{i,j}^{(3)} = \frac{1}{\Delta t^3} \sum_{r=-1}^{2} \gamma_{2,r}^{3,0} f_{i,j}^{3,n+r};$$

Step 13: Compute $f^{(3)}_{i+\frac{1}{2}}$ adopting the interpolatory formula on stencil $S^2_{i+\frac{1}{2}}$ as:

$$f^{(3)}_{i+\frac{1}{2}} = \sum_{j=-1}^{2} \gamma^{0,\frac{1}{2}}_{2,j} f^{(3)}_{i,j};$$

Step 14: Reconstruct $F^2_{i+\frac{1}{2}}$ from (3.1.2) as:

$$F^2_{i+\frac{1}{2}} = f^{(0)}_{i+\frac{1}{2}} + \Delta t f^{(1)}_{i+\frac{1}{2}} + \frac{\Delta t^2}{2} f^{(2)}_{i+\frac{1}{2}} + \frac{\Delta t^3}{6} f^{(3)}_{i+\frac{1}{2}}$$

3.2 Adptive Compact Approximate Taylor Method

The shock-capturing methods are a class of numerical techniques for computing inviscid flows with shock waves. Computation of flow through shock waves is an extremely difficult task because such flows results in sharp, discontinuous changes in flow variables pressure, density, temperature, and velocity across the shock.

From an historical point of view, shock-capturing methods can be classified into two general categories: classical methods and modern shock capturing methods (also called high-resolution schemes). Modern shock-capturing methods are generally upwind based in contrast to classical symmetric or central discretization. Upwind-type differencing schemes attemp to discretize hyperbolic partial differential equation by using differencing biased in the direction determined by the sign of the characteristic speeds. On the other hand, symmetric or central schemes do not consider any information about the wave propagation in the discretization.

No matter what type of shock-capturing scheme is used, a stable calculation in presence of shock wave requires a certain amount of numerical dissipation, in order to avoid the formation of spurious numerical oscillation as could be observed in section 3.1.2 [55]. In the case of classical shock capturing methods, numerical dissipation terms are usually linear and the same amount is uniformly applied to all grid points. Classical shock-capturing methods only exhibit accurate results in the case of smooth and weak-shock solution, but when strong shock waves are present, non-linear instabilities and oscillations can arise across discontinuities. Modern shock-capturing methods have, however, a non-linear numerical dissipation, with an automatic feedback mechanism which adjust the amount of the dissipation in any cell of the mesh, in accord with the gradients in the solution.

3.2.1 Flux-Limiter schemes

As part of the classical shock-capturing family, the Flux-Limiter methods adopt a numerical techniques to combine a low order method with an high order method and make the solution a total variation diminishing solution [72, 124]. The main idea behind the construction of flux limiter schemes is to limit the spatial derivatives to realistic values. They are used in high resolution schemes for solving problems described by partial differential equations and only come into operation when sharp wave fronts are present. From smoothly changing waves, the flux limiters do not operate and the spatial derivative can be represented by higher order approximation without introducing of spurious oscillation.

Let us consider the scalar conservation law (2.1.7). In particular, we will focus on the 1D scalar semi-discrete scheme below:

$$\frac{du_i}{dt} + \frac{1}{\Delta x}\Big(f(u_{i+\frac{1}{2}}) - f(u_{i-\frac{1}{2}})\Big) = 0, \tag{3.2.1}$$

where, $f(u_{i\pm\frac{1}{2}})$, represent edge fluxes for the $i-$th cell. If these edge fluxes should be written by a low and an high order reconstruction, then a flux limiter scheme can switch between these reconstructions depending upon the gradients close to the particular cell, as follow,

$$F^*_{i+\frac{1}{2}} = \varphi_{i+\frac{1}{2}} F^1_{i+\frac{1}{2}} + (1 - \varphi_{i+\frac{1}{2}}) F^{lo}_{i+\frac{1}{2}}, \tag{3.2.2}$$

in which, $F^1_{i+\frac{1}{2}}$ is given by (3.1.5); $F^{lo}_{i+\frac{1}{2}}$ is a first-order robust numerical flux; and $\varphi_{i+\frac{1}{2}}$ is a standard flux limiter function, see [70, 83, 84, 119, 120].

In general, a classical flux limiter function depends on the gradients of the solution, i.e.

$$\varphi_{i+\frac{1}{2}} = \varphi(r_{i+\frac{1}{2}}),$$

where

$$r_{i+\frac{1}{2}} = \frac{\Delta upw}{\Delta loc} = \begin{cases} r^-_{i+\frac{1}{2}} := \frac{u^n_i - u^n_{i-1}}{u^n_{i+1} - u^n_i} & \text{if } a_{i+\frac{1}{2}} > 0, \\ r^+_{i+\frac{1}{2}} := \frac{u^n_{i+2} - u^n_{i+1}}{u^n_{i+1} - u^n_i} & \text{if } a_{i+\frac{1}{2}} \leq 0; \end{cases} \tag{3.2.3}$$

and $a_{i+\frac{1}{2}}$ is an estimate of the wave speed, for instance Roe's intermediate speed:

$$a_{i+\frac{1}{2}} = \begin{cases} \frac{f(u_{i+1}^n)-f(u_i^n)}{u_{i+1}^n-u_i^n} & \text{if } u_i^n \neq u_{i+1}^n, \\ f'(u_i^n) & \text{otherwise.} \end{cases}$$

An alternative to avoids the computation of an intermediate speed was introduced in [120]: it consists in defining

$$\varphi_{i+\frac{1}{2}} = \min\Big(\varphi(r^-_{i+\frac{1}{2}}), \varphi(r^+_{i+\frac{1}{2}})\Big).$$

Remark 3.2.1 *Note that, the flux limiter functions do not depend on the method used, even if, it is finite difference or finite volume scheme. They reckon only the numerical data in the local stencil.*

Remark 3.2.2 *Note that, even if the numerical schemes are 3-points methods, the flux limiter functions need a 4-points stencil to be computed. This is due to the fact that is impossible distinguish between a critical point or discontinuity adopting a 3-points stencil.*

For the system of conservation law (2.3.1), the expression of the flux limiter function is similar to the scalar case. Indeed, following the Toro's idea [120], computed the flux limiter functions for each variable of the system $\varphi^\ell_{i+\frac{1}{2}}$ for all $\ell = 1, \ldots, m$ we set:

$$\varphi_{i+\frac{1}{2}} = \min_{\ell=1,\ldots,m} \varphi^\ell_{i+\frac{1}{2}}. \tag{3.2.4}$$

As we have seen, given $r_{i+\frac{1}{2}}$, there are several ways to define the operator $\varphi(r_{i+\frac{1}{2}})$, we will focus on two different flux limiter functions: Minmod and SuperBee, respectively,

$$\varphi^{\text{minmod}}_{i+\frac{1}{2}} = \begin{cases} 1 & \text{if } r_{i+\frac{1}{2}} > 1 \\ r & \text{if } 0 < r_{i+\frac{1}{2}} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.5}$$

$$\varphi^{\text{superbee}}_{i+\frac{1}{2}} = \begin{cases} 2 & \text{if } r_{i+\frac{1}{2}} > 2 \\ r & \text{if } 1 < r_{i+\frac{1}{2}} \leq 2 \\ 1 & \text{if } \frac{1}{2} < r_{i+\frac{1}{2}} \leq 1 \\ 2r & \text{if } 0 < r_{i+\frac{1}{2}} \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.2.6}$$

In order to show the difference between Minmod and SuperBee function two simple numeri-

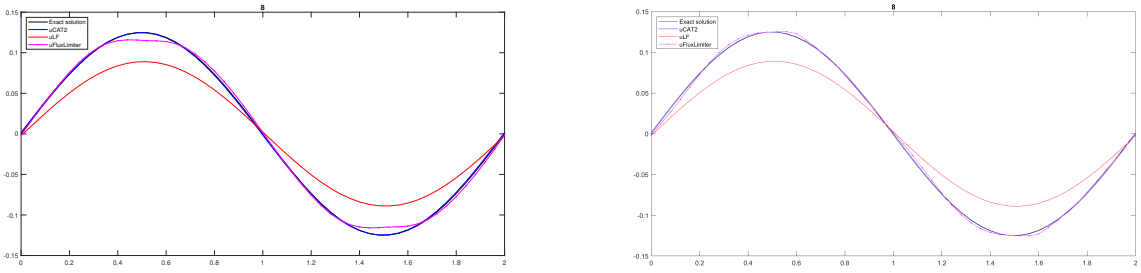

Figure 3.2.1: Transport equation with smooth initial condition. Numerical solutions at time $t = 8$ obtained with CAT2, Lax-Friedrichs and Flux Limiter methods using a 50−points mesh and CFL= 0.9. *left* the flux limiter solutions with Minmod function (3.2.5); *right* the flux limiter solutions with SuperBee function (3.2.6).

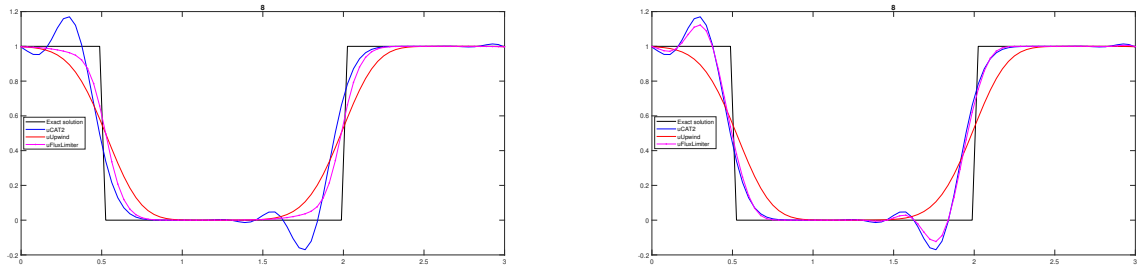

Figure 3.2.2: Transport equation with no-smooth initial condition. Numerical solutions at time $t = 8$ obtained with CAT2, Upwind and Flux Limiter methods using a 50−points mesh and CFL= 0.9. *left* the flux limiter solutions with Minmod function (3.2.5); *right* the flux limiter solutions with SuperBee function (3.2.6).

cal examples are tested. For this reason, let us consider the scalar transport equation (2.1.2) and two different initial conditions:

1. smooth initial condition

$$u_0(x) = \frac{1}{8}\sin(2\pi x)$$

   defined on $[0, 2]$, adopting a 50−points mesh, CFL= 0.9, periodic boundary conditions and final time $t = 8$;

2. double jump initial condition

$$u_0(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1.5 \\ 0 & \text{otherwise,} \end{cases}$$

   defined on $[0, 3]$, adopting a 80−points mesh, CFL= 0.9, periodic boundary conditions and final time $t = 8$;

Figure 3.2.1 and Figure 3.2.2 show the behaviour of the Minmod and SuperBee flux limiter functions. In particular, the Minmod function gives better results close to a discontinuity but is more diffusive near a critical point. Instead the SuperBee function gives better results close a critical point but introduce some spurious numerical oscillations near a discontinuity. These behaviours should be reduced using different flux limiter functions but they cannot be eliminated [72]. To avoid these problems an high order smoothness indicators should be used and it will be focus of next section.

3.2.2 High order smoothness indicators

In this section a new family of local smoothness indicators $\psi^p_{i+\frac{1}{2}}$, $p \geq 2$, for scalar conservation laws and their properties will be introduced.

In particular, fixed the nodal approximations $f_i = f(x_i)$ of a function f at the stencil $S^p_{i+\frac{1}{2}}$, $p \geq 2$, centered at $x_{i+\frac{1}{2}}$, first define the lateral weights:

$$I_{p,L} := \sum_{j=-p+1}^{-1} (f_{i+1+j} - f_{i+j})^2 + \varepsilon, \quad I_{p,R} := \sum_{j=1}^{p-1} (f_{i+1+j} - f_{i+j})^2 + \varepsilon, \tag{3.2.7}$$

where ε is a small quantity that is added to prevent that the lateral weights vanish when the function is constant. Next, compute the quantity:

$$I_p := \frac{I_{p,L} I_{p,R}}{I_{p,L} + I_{p,R}}. \tag{3.2.8}$$

Finally, define the smoothness indicator of the stencil of $S^p_{i+\frac{1}{2}}$ by

$$\psi^p_{i+1/2} := \left(\frac{I_p}{I_p + \tau_p} \right)^2, \tag{3.2.9}$$

where

$$\tau_p := \left(\Delta^{2p-1}_{i-p+1} f \right)^2. \tag{3.2.10}$$

Here, $\Delta^{2p-1}_{i-p+1} f$ denotes the undivided difference of $\{f_{i-p+1}, \ldots, f_{i+p}\}$:

$$\Delta^{2p-1}_{i-p+1} f = (2p-1)! \sum_{j=-p+1}^{p} \gamma^{2p-1,1/2}_{p,j} \, f_{i+j}. \tag{3.2.11}$$

Before going into technical details, let us give a motivation of this choice. In fact, if data in the stencil $S^p_{i+\frac{1}{2}}$ are smooth, then

$$I_{p,L} = O(\Delta x^2), \quad I_{p,R} = O(\Delta x^2), \quad \tau_p = O(\Delta x^{4p}).$$

Since

$$\frac{1}{I_p} = \frac{1}{I_{p,L}} + \frac{1}{I_{p,R}}$$

then $I_p = O(\Delta x^2)$ and thus

$$\psi^p_{i+1/2} = \frac{I_p}{I_p + \tau_p} = \frac{O(\Delta x^2)}{O(\Delta x^2) + O(\Delta x^{4p})},$$

so that $\psi^p_{i+\frac{1}{2}}$ is expected to be close to 1. On the other hand, if there is an isolated discontinuity in the stencil then

$$\tau_p = O(1)$$

and, one between left or right lateral weights presents the discontinuity implying that:

$$I_{p,L} = O(1), \quad I_{p,R} = O(\Delta x^2)$$

or

$$I_{p,L} = O(\Delta x), \quad I_{p,R} = O(1).$$

In both cases $I_p = O(\Delta x^2)$ and thus:

$$\psi^p_{i+1/2} = \frac{I_p}{I_p + \tau_p} = \frac{O(\Delta x^2)}{O(\Delta x^2) + O(1)},$$

so that $\psi^p_{i+\frac{1}{2}}$ is expected to be close to 0. Nevertheless, in general it is not true that

$$\frac{O(\Delta x^2)}{O(\Delta x^2) + O(\Delta x^{4p})} \approx 1, \quad \frac{O(\Delta x^2)}{O(\Delta x^2) + O(1)} \approx 0,$$

and a careful analysis is required. In the case of smooth data, special care has to be taken if there is a critical point in the stencil, since in this case the order of I_p depends on the order of the critical point, what can prevent the smoothness indicator to be close to 1, as it will be seen in Propositions 3.2.1-3.2.3 below. The following definition is assumed in these results: a point x is said to be a critical point of f of order n if $f^{(j)}(x) = 0$, $j = 1, \ldots, n$ and

$f^{(n+1)} \neq 0$.

Before analysing the smoothness indicators, let us introduce some definitions and notation, taken from [2]: we refer to Section 2.1 of this reference for further details.

Given $\alpha \in \mathbb{R}^+$ and $f : (0, h^*) \mapsto \mathbb{R}$ with $h^* \in (0, \infty]$, the notation $f(h) = \mathcal{O}(h^\alpha)$ means, as usual, that

$$\limsup_{h \to 0^+} \left| \frac{f(h)}{h^\alpha} \right| < +\infty,$$

and the notation $f(h) = \overline{\mathcal{O}}(h^\alpha)$ means that

$$\limsup_{h \to 0^+} \left| \frac{f(h)}{h^\alpha} \right| < +\infty \quad \text{and} \quad \liminf_{h \to 0^+} \left| \frac{f(h)}{h^\alpha} \right| > 0.$$

If $f, g : (0, h^*) \mapsto \mathbb{R}$ and α, β are two positive real numbers, the following relations hold:

$$\begin{aligned}
&f(h) = \mathcal{O}(h^\alpha), \quad g(h) = \mathcal{O}(h^\beta) \implies f(h)g(h) = \mathcal{O}(h^{\alpha+\beta});\\
&f(h) = \overline{\mathcal{O}}(h^\alpha), \quad g(h) = \overline{\mathcal{O}}(h^\beta) \implies f(h)g(h) = \overline{\mathcal{O}}(h^{\alpha+\beta});\\
&f > 0, f(h) = \overline{\mathcal{O}}(h^\alpha) \implies f(h)^{-1} = \overline{\mathcal{O}}(h^{1/\alpha}).
\end{aligned}$$

Lemma 3.2.1 *Assume that a function $\varphi \in \mathcal{C}^{n+2}$ satisfies $\varphi^{(k)}(0) = 0$ for $k = 1, \ldots, n$ and $\varphi^{(n+1)}(0) \neq 0$. Then $\varphi(h) = \overline{\mathcal{O}}(h^{n+1})$*

Proof. The Taylor expansion of φ truncated at order $n+1$ is so set:

$$\varphi(h) = \frac{\varphi^{(n+1)(0)}}{(n+1)!} h^{n+1} + \mathcal{O}(h^{n+2}),$$

which implies

$$\lim_{h \to 0} \frac{\varphi(h)}{h^{n+1}} = \frac{\varphi^{(n+1)(0)}}{(n+1)!} \neq 0.$$

Then, $\varphi(h) = \overline{\mathcal{O}}(h^{n+1})$.

■

Lemma 3.2.2 *Let $c, d, z \in \mathbb{R}$. Assume that*

$$\begin{cases}
f^{(j)}(z) = 0 \text{ for } j = 1, \ldots, k, \quad f^{(k+1)}(z) \neq 0, \text{ and } f \in \mathcal{C}^{k+2} & \text{if } c + d \neq 0;\\
f^{(2j-1)}(z) = 0 \text{ for } j = 1, \ldots, n, \quad f^{(2n+1)}(z) \neq 0, \text{ and } f \in \mathcal{C}^{2n+2} & \text{if } c + d = 0.
\end{cases}$$

Then

$$f(z+dh) - f(z+ch) = \overline{\mathcal{O}}(h^s),$$

where

$$s = \begin{cases} k+1 & \text{if } c+d \neq 0; \\ 2n+1 & \text{if } c+d = 0. \end{cases}$$

From this lemma, whose proof follows from Lemma 3.2.1 more details in [2], one can deduce that, given the values $f_j = f(x_j)$, $j = i-p+1, \ldots, i+p$, of a smooth enough function f in the stencil $S^p_{i+\frac{1}{2}}$, the following estimates hold:

$$f_{j+1} - f_j = \mathcal{O}(h), \quad j = i-p+1, \ldots, i+p-1$$

if the stencil does not contain any critical point of f;

$$f_{j+1} - f_j = \overline{\mathcal{O}}(h^{k+1}), \quad j = i-p+1, \ldots, i+p-1, \tag{3.2.12}$$

if the stencil contains a critical point x^* of even order k or a critical point of odd order that is not located at the center of any sub-interval of the stencil.

Finally, if there exists i_0 such that $x^* = 0.5(x_{i_0} + x_{i_0+1})$, x^* is a critical point of odd order, then (3.2.12) holds for every $j \neq i_0$ and

$$f_{i_0+1} - f_{i_0} = \overline{\mathcal{O}}(h^{2n+1}) \tag{3.2.13}$$

where $2n+1$ is the first odd number such that

$$f^{(2n+1)}(x^*) \neq 0.$$

Let us analyze the behavior of the smoothness indicators (3.2.9) assuming that $\varepsilon = 0$ (the role of ε is only relevant for the implementation of the method not for the analysis).

Proposition 3.2.1 *Let $f_j = f(x_j)$, $j = i-p+1, \ldots, i+p$ be the values of a function f in the stencil $S^p_{i+\frac{1}{2}}$, with $p > 2$. The following estimates hold:*

$$\psi^p_{i+1/2} = \begin{cases} 1 - \mathcal{O}(\Delta x^{4(p-1)-2k}) & \text{if } f \in \mathcal{C}^{\max(2p-1,k+2)}; \\ \overline{\mathcal{O}}(\Delta x^{2(k+1)}) & \text{if } f \text{ is piecewise } \mathcal{C}^{k+2} \text{ and } S^p_{i+\frac{1}{2}} \text{ contains an isolated jump discontinuity of } f \end{cases}$$

where $k = 0$ if there is no critical point of f in S_p or k equal to the order of the critical point if there is one.

Proof. If $f \in C^{2p-1}$ there exists ξ such that

$$\Delta^{2p-1}_{i-p+1} f = (2p-1)! f^{(2p-1)}(\xi) \Delta x^{2p-1},$$

and thus

$$\Delta^{2p-1}_{i-p+1} f = \mathcal{O}(\Delta x^{2p-1}),$$

what implies

$$\tau_p = \mathcal{O}(\Delta x^{4p-2}).$$

On the other hand, if $S^p_{i+\frac{1}{2}}$ contains an isolated jump discontinuity, then

$$\Delta^{2p-1}_{i-p+1} f = \mathcal{O}(1),$$

and thus

$$\tau_p = \overline{\mathcal{O}}(1).$$

From the discussion above, the estimate

$$f_{j+1} - f_j = \overline{\mathcal{O}}(\Delta x^{k+1}),$$

holds for every $j \in i-p+1, \ldots, i+p-1$ with the exception of at most one index i_0, in which the order is higher.

Nevertheless, since both $I_{p,L}$ and $I_{p,R}$ are the sum of at least two terms of the form $(f_{j+1} - f_j)^2$, we can conclude that

$$I_{p,L} = \overline{\mathcal{O}}(\Delta x^{2+2k}), \quad I_{p,R} = \overline{\mathcal{O}}(\Delta x^{2+2k}).$$

Hence:

$$I_p = \frac{I_{p,L} I_{p,R}}{I_{p,L} + I_{p,R}} = \frac{\overline{\mathcal{O}}(\Delta x^{2+2k}) \overline{\mathcal{O}}(\Delta x^{2+2k})}{\overline{\mathcal{O}}(\Delta x^{2+2k}) + \overline{\mathcal{O}}(\Delta x^{2+2k})} = \frac{\overline{\mathcal{O}}(\Delta x^{4+4k})}{\overline{\mathcal{O}}(\Delta x^{2+2k})} = \overline{\mathcal{O}}(\Delta x^{2+2k}).$$

Now, if $S^p_{i+\frac{1}{2}}$ contains a discontinuity, then, by construction, there exists a side $\alpha \in \{\mathrm{L}, \mathrm{R}\}$ such that $I_{p,\alpha} = \overline{\mathcal{O}}(1)$ (the side that contains the discontinuity) while the other side, $\beta \in$

$\{\mathrm{L},\mathrm{R}\}\setminus\{\alpha\}$, satisfies $I_{p,\beta}=\overline{\mathcal{O}}(\Delta x^{2+2k})$. Therefore

$$I_p=\frac{I_{p,L}I_{p,R}}{I_{p,L}+I_{p,R}}=\frac{I_{p,\alpha}I_{p,\beta}}{I_{p,\alpha}+I_{p,\beta}}=\frac{\overline{\mathcal{O}}(1)\overline{\mathcal{O}}(\Delta x^{2+2k})}{\overline{\mathcal{O}}(1)+\overline{\mathcal{O}}(\Delta x^{2+2k})}=\frac{\overline{\mathcal{O}}(\Delta x^{2+2k})}{\overline{\mathcal{O}}(1)}=\overline{\mathcal{O}}(\Delta x^{2+2k}).$$

Combining the above results, we have that, if f is smooth:

$$\psi^p_{i+1/2}=\frac{I_p}{I_p+\tau_p}=\frac{1}{1+\dfrac{\tau_p}{I_p}}=\frac{1}{1+\dfrac{\mathcal{O}(\Delta x^{4p-2})}{\overline{\mathcal{O}}(\Delta x^{2+2k})}}=\frac{1}{1+\mathcal{O}(\Delta x^{4(p-1)-2k})}=1-\mathcal{O}(\Delta x^{4(p-1)-2k}).$$

On the other hand, if $S^p_{i+\frac{1}{2}}$ contains a discontinuity, then

$$\psi^p_{i+1/2}=\frac{I_p}{I_p+\tau_p}=\frac{1}{1+\dfrac{\tau_p}{I_p}}=\frac{1}{1+\dfrac{\overline{\mathcal{O}}(1)}{\overline{\mathcal{O}}(\Delta x^{2+2k})}}=\frac{1}{1+\overline{\mathcal{O}}(\Delta x^{-2(k+1)})}=\overline{\mathcal{O}}(\Delta x^{2(k+1)}),$$

which finishes the proof. ■

Observe that the indicator $\psi^p_{i+\frac{1}{2}}$ is able to detect smoothness in the presence of a critical point whose order is lower than $2(p-1)$.

In the case $p=2$ similar arguments lead to the following estimates:

Proposition 3.2.2 *Let $f_j=f(x_j)$, $j=i-1,\dots,i+2$ be the values of a function f in the stencil $S^2_{i+\frac{1}{2}}$. The following estimates hold:*

$$\psi^2_{i+1/2}=\begin{cases}1-\mathcal{O}(\Delta x^{4-2k}) & \text{if } f\in\mathcal{C}^3;\\ \overline{\mathcal{O}}(\Delta x^{2(k+1)}) & \text{if } f \text{ is piecewise } \mathcal{C}^{k+2} \text{ and } S^2_{i+\frac{1}{2}} \text{ contains an isolated jump discontinuity of } f;\end{cases}$$

where $k=0$ if there is no critical point of f in $S^2_{i+\frac{1}{2}}$ and $k=1$ if there is a critical point x^* of order 1 such that $f^{(3)}(x^*)\neq 0$ or such that $x^*\neq 0.5(x_j+x_{j+1})$ for $j=i-1,i+1$.

Nevertheless, the estimate cannot be proved when $S^2_{i+\frac{1}{2}}$ includes a critical point of order 1 located at $0.5(x_{i-1}+x_i)$ or $0.5(x_{i+1}+x_{i+2})$ and such that $f^{(3)}(x^*)\neq 0$: the argument in the proof of Proposition 3.2.1 cannot be used since there is only one term in the definition of the local weights. This is not a limitation in many applications, since this situation is very specific and, even if it happens, unless there is a discontinuity close to the critical point, smoothness will be detected by at least one of the indicators $\psi^p_{i+\frac{1}{2}}$ with $p>2$ so that the stencil $S^p_{i+\frac{1}{2}}$ will be used to update the solution. In any case, the smoothness indicator for

$p = 2$ can be modified to properly handle these situations as follows: compute the couple of lateral weights:

$$I^1_{2,L} := (f_i - f_{i-1})^2 + \varepsilon, \quad I^1_{2,R} := (f_{i+1} - f_i)^2 + (f_{i+2} - f_{i+1})^2 + \varepsilon, \tag{3.2.14}$$

$$I^2_{2,L} := (f_i - f_{i-1})^2 + (f_{i+1} - f_i)^2 + \varepsilon, \quad I^2_{2,R} := (f_{i+2} - f_{i+1})^2 + \varepsilon. \tag{3.2.15}$$

Next, compute:

$$I^j_2 := \frac{I^j_{2,L} I^j_{2,R}}{I^j_{2,L} + I^j_{2,R}}, \quad j = 1, 2. \tag{3.2.16}$$

and then, the smoothness indicator of the stencil $S^2_{i+\frac{1}{2}}$ is given by

$$\widetilde{\psi}^2_{i+1/2} := \max\left(\frac{I^1_2}{I^1_2 + \tau_2}, \frac{I^2_2}{I^2_2 + \tau_2}\right). \tag{3.2.17}$$

The following estimate can be then proved:

Proposition 3.2.3 *Let $f_j = f(x_j)$, $j = i-1, \ldots, i+2$ be the values of a function f in the stencil $S^2_{i+\frac{1}{2}}$. The following estimates hold:*

$$\widetilde{\psi}^2_{i+1/2} = \begin{cases} 1 - \mathcal{O}(\Delta x^{4-2k}) & \text{if } f \in \mathcal{C}^3; \\ \overline{\mathcal{O}}(\Delta x^{2(k+1)}) & \text{if } f \text{ is piecewise } \mathcal{C}^{k+2} \text{ and } S^2_{i+\frac{1}{2}} \text{ contains an isolated jump discontinuity of } f; \end{cases}$$

where $k = 0$ if there is no critical points of f in $S^2_{i+\frac{1}{2}}$ or $k = 1$ if there is a critical point x^* or order 1.

Proof. The arguments of the proof of Proposition 3.2.1 are used again. The difference comes from the case in which there is a critical point of order 1 located at at $0.5(x_{i-1} + x_i)$ or $0.5(x_{i+1} + x_{i+2})$ and such that $f^{(3)}(x^*) = 0$. In this case, there exists $j \in \{1, 2\}$ (the one in which the sub-interval with the critical point and the central sub-interval are considered together in the same lateral weight) such that

$$\frac{I^j_2}{I^j_2 + \tau_2} = 1 - \mathcal{O}(\Delta x^2).$$

Using this estimate the proof is concluded as in Proposition 3.2.1 ■

Remark 3.2.3 *The smoothness indicators* (3.2.9) *and* (3.2.17) *have the following homoth-*

etic invariance property: given a function f and two positive numbers α, β, define

$$g(x) = \alpha f(\beta x).$$

Then the smoothness indicator of f at the stencil $S^p_{i+\frac{1}{2}}$ centered at $x_{i+\frac{1}{2}}$ in a mesh with step Δx is equal to the smoothness indicator of g at the stencil $S^p_{i+\frac{1}{2}}$ centered at $\beta x_{i+\frac{1}{2}}$ in a mesh with step $\beta\Delta x$. This property is very important in order to construct smoothness indicators whose behaviour do not depend on Δx and scaling factors of f.

3.2.3 ACAT2P

As we have seen on previous section, the Compact Approximate Taylor (CAT) schemes introduce spurious oscillations close to a discontinuity of the solution under the usual CFL condition, as it happens for the Lax-Wendroff method: see [15, 26, 55, 56]. A shock-capturing technique, based on a family of high-order smoothness indicators, is considered here to prevent this behaviour, developed in [14]. The idea is as follows: once the approximations at time t^n have been computed, the candidate stencils to compute $F^A_{i+\frac{1}{2}}$ are

$$S^p_{i+\frac{1}{2}} = \{x_{i-p+1}, \ldots, x_{i+p}\}, \quad p = 1, \ldots, P.$$

The selected stencil is the one with maximal length among those in which the solution at time t^n is smooth, according to smoothness indicators (3.2.9) $\psi^p_{i+\frac{1}{2}}$ for $p = 1, \ldots, P$ introduced on previous section. If a discontinuity is detected in the stencil $S^1_{i+\frac{1}{2}}$ a robust first-order numerical method is used.

In order to select the stencil, the smoothness indicators (3.2.9) $\psi^p_{i+\frac{1}{2}}$, $p = 1, \ldots, P$ are computed such that:

$$\psi^p_{i+\frac{1}{2}} \approx \begin{cases} 1 & \text{if } \{u^n_i\} \text{ is 'smooth' in } S^p_{i+\frac{1}{2}}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.18}$$

Define now:

$$\mathcal{A} = \{p \in \{1, \ldots, P\} \text{ such that } \psi^{\mathrm{p}}_{\mathrm{i}+\frac{1}{2}} \approx 1\}.$$

The idea is to define:

$$F^A_{i+\frac{1}{2}} = \begin{cases} F^{lo}_{i+\frac{1}{2}} & \text{if } \mathcal{A} = \emptyset; \\ F^{p_s}_{i+\frac{1}{2}} & \text{where } p_s = \max(\mathcal{A}) \text{ otherwise;} \end{cases}$$

where $F^{p_s}_{i+\frac{1}{2}}$ is the numerical flux of CAT2p_s and $F^{lo}_{i+\frac{1}{2}}$ is a robust first order numerical flux. Nevertheless, it is not possible to determine if the solution is smooth or not in the stencil $S^1_{i+\frac{1}{2}}$ where only two values u^n_i, u^n_{i+1} are available. Therefore in practice, will be defined:

$$\mathcal{A} = \{p \in \{2, \dots, P\} \text{ such that } \psi^{\mathrm{p}}_{\mathrm{i}+\frac{1}{2}} \approx 1\}. \tag{3.2.19}$$

and then:

$$F^A_{i+\frac{1}{2}} = \begin{cases} F^*_{i+\frac{1}{2}} & \text{if } \mathcal{A} = \emptyset; \\ F^{p_s}_{i+\frac{1}{2}} & \text{where } p_s = \max(\mathcal{A}) \text{ otherwise;} \end{cases} \tag{3.2.20}$$

where $F^*_{i+\frac{1}{2}}$ is the numerical flux reconstruction obtained through the flux-limiter scheme that combine a first robust method with CAT2 as (3.2.2) (that uses the stencil $S^1_{i+\frac{1}{2}}$ as well).

The expression of the Adaptive Compact Approximate Taylor Method (ACAT2P) of maximal order $2P$ for a scalar conservation law is given by:

$$u^{n+1}_i = u^n_i + \frac{\Delta t}{\Delta x}\left(F^A_{i-\frac{1}{2}} - F^A_{i+\frac{1}{2}}\right). \tag{3.2.21}$$

The numerical fluxes $F^A_{i+\frac{1}{2}}$ are defined by (3.2.19)-(3.2.20). For $p = 2$, in order to avoid problems to select correctly the smoothness of the solution, in presence of critical point, (3.2.9) can be replaced by (3.2.17).

Observe that, by definition, $F^A_{i+\frac{1}{2}}$ reduces to:

- a first order flux if $\psi^1_{i+\frac{1}{2}} = 0$ and $\psi^p_{i+\frac{1}{2}} = 0$ for all $p = 2, \dots, P$;
- a second order flux if $\psi^1_{i+\frac{1}{2}} = 1$ and $\psi^p_{i+\frac{1}{2}} \approx 0$ for all $p = 2, \dots, P$;
- $2p_s$-order flux if $\psi^{p_s}_{i+\frac{1}{2}} \approx 1$.

Furthermore, if $p_s = P$, then ACAT2P coincides with CAT2P which has $2P$-order accuracy and is L^2-stable under CFL≤ 1.

Proposition 3.2.4 *The local accuracy of the method close to a critical point is always $2P$ with the only exception of critical points of order $2P-2$: in that case, the order of accuracy*

will be reduced by one.

Proof. Let us suppose that f is smooth and has an isolated critical point x^* of order k in $S^1_{i+\frac{1}{2}} = \{x_i, x_{i+1}\}$. Then:

- If $k < 2(P-1)$ the smoothness indicator $\psi^P_{i+\frac{1}{2}}$ is close to one and the maximum allowed stencil S_P is used, so that the local accuracy of the method is $2P$.

- If $k > 2(P-1)$ then all the smoothness indicators fail, so that the first order robust numerical method will be used. Nevertheless in this case, $f^{(j)}(x^*) = 0$ for $j = 1, \ldots, 2P-1$ so that, when the local error of the first order method is estimated through Taylor expansions, only terms of order $O(\Delta x^{2P})$ or bigger will remain. Therefore, in this case the local accuracy of the method is again $2P$.

- If $k = 2(P-1)$ again the smoothness indicators fail and the first order robust numerical method will be used. Since in this case, $f^{(j)}(x^*) = 0$ for $j = 1, \ldots, 2P-2$ the local error of the first order method is of order $2P-1$. ■

This order reduction should be avoided by introducing optimal smoothness indicators in the spirit of [2, 3].

High order smoothness indicators for systems of conservation laws

For systems of conservation laws (2.3.1) with d equations, the expression of the ACAT2P method is the same as in the scalar case: the only difference is the computation of the smoothness indicators. In the case of systems, smoothness indicators are first computed for every variable:

$$\psi^{\ell,p}_{i+\frac{1}{2}}, \quad p = 1, \ldots, P,$$

where

- $\psi^{\ell,1}_{i+\frac{1}{2}} = \varphi^{\ell}_{i+\frac{1}{2}}$ is an usual the flux limiter (3.2.4) computed following the Toro's idea [120] for each $\ell-$th component of the numerical solutions $\{u^{\ell,n}_i\}$;

- $\psi^{\ell,p}_{i+\frac{1}{2}}$, $p > 2$ is obtained by applying the smoothness indicator (3.2.9) to the $\ell-$th component of the numerical solutions $\{u^{\ell,n}_i\}$;

- $\psi^{\ell,2}_{i+\frac{1}{2}}$ is obtained by applying the smoothness indicator (3.2.9) or (3.2.17) to the $\ell-$th component of the numerical solutions $\{u^{\ell,n}_i\}$.

Once these scalar smoothness indicators have been computed, we define

$$\psi^p_{i+\frac{1}{2}} = \min_{\ell=1,\dots,d} \psi^{\ell,p}_{i+\frac{1}{2}},$$

so that the selected stencil is the one of maximal length among those in which all the variables are smooth.

3.2.4 Numerical experiments

In this section we focus on the behaviour of the ACAT2P procedure applied to several 1D problems: the linear transport equation, Burgers equation, and the Euler equation for gas dynamic. As flux limiter function, for the Flux-Limiter techniques (ACAT2), the Super Bee flux limiter [105] is used; and the smoothness indicators (3.2.9), for ACAT2P, are used for $p \geq 2$: no loss of precision for first order critical points has been observed in any of the test problems considered here due to the use of $\psi^1_{i+1/2}$. Fornberg's algorithm [41, 42] is used to compute iteratively the coefficients of the numerical differentiation formulas. ACAT2P, for $p = 2, 4, 6$, methods will be compared with the Lax-Friedrichs (LF), HLL first order schemes and with WENO($2p+1$) finite difference methods based on the Lax-Friedrichs splitting in Chapter 1 (see [110]) combined with SSPRK3 in Chapter 1 (see [48]) for the time discretization. The number of points of their stencils and the relative theoretical order in 1D are recalled in Table 3.2. Since ACAT2P reduces to CAT2P in smooth region and the order of accuracy of the latter has been checked in [15], no other test order for the systems of conservation laws will be considered here.

Method	Stencil	Order
LF	3	1 in space and time
HLL	3	1 in space and time
FL-CAT2 or ACAT2	3	2 in space and time
ACAT2P	$2P+1$	$2P$ in space and time
WENO($2p+1$)-RK3	$2p+1$	$2p+1$ in space and 3 in time

Table 3.2: Numerical methods: number of points of the stencil and order of accuracy for 1D problems.

In practice, on equation (3.2.18) we impose $\psi^p_{i+\frac{1}{2}} \approx 1$. The following criterion to check the proximity of $\psi^p_{i+\frac{1}{2}}$ to 1 has been implemented in order to define the admissible set of

indices (3.2.19) as:

$$\mathcal{A} = \{p \in \{2, \dots, P\} \; s.t. \; \psi^p_{i+\frac{1}{2}} \geq 0.9\}. \tag{3.2.22}$$

1D scalar transport equation

Let us consider the linear scalar transport equation

$$u_t + u_x = 0. \tag{3.2.23}$$

We solve it with different type of initial condition including smooth and no-smooth conditions.

Test 1: We consider the transport equation (3.2.23) with smooth initial condition:

$$u_0(x) = \frac{1}{2}\sin(\pi x). \tag{3.2.24}$$

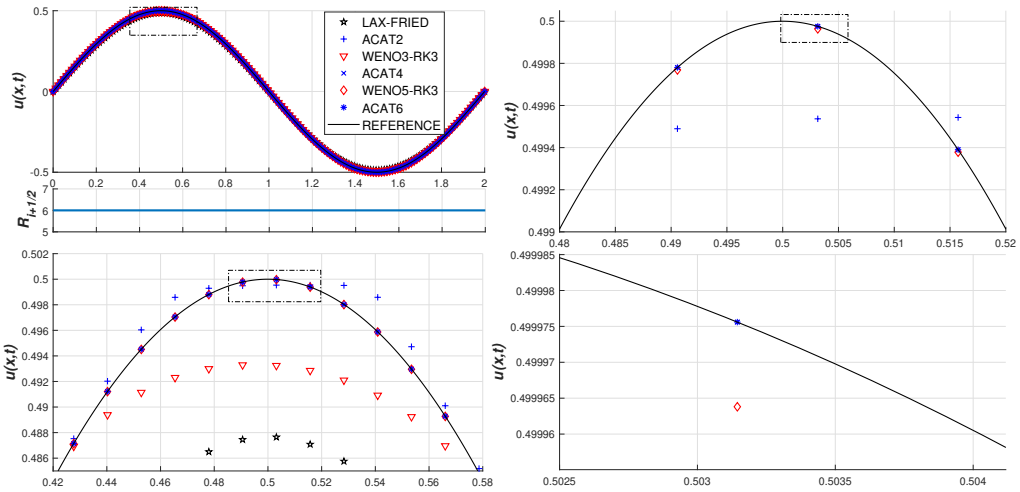

Figure 3.2.3: Transport equation with initial condition (3.2.24). Numerical solutions at $t = 3$: general view (*left-top*); order of accuracy for ACAT6 (*sub-frame*); consecutive zooms close to the local maximum (*left-bottom*, *right-top* and *right-bottom*).

We solve numerically this problem (3.2.23) in the interval $[0, 2]$, using 160 mesh points, CFL= 0.9, and periodic boundary conditions. Figure 3.2.3 and 3.2.4 show the numerical solutions at time $t = 3$ and $t = 40$ respectively. Zooms of an interest area are included, in which the loss of accuracy with time for the lower order methods can be clearly seen. As it can be observed, the numerical solutions of ACAT4 and ACAT6 match the exact solution at both times while ACAT2 is more diffusive near the critical points. This loss of

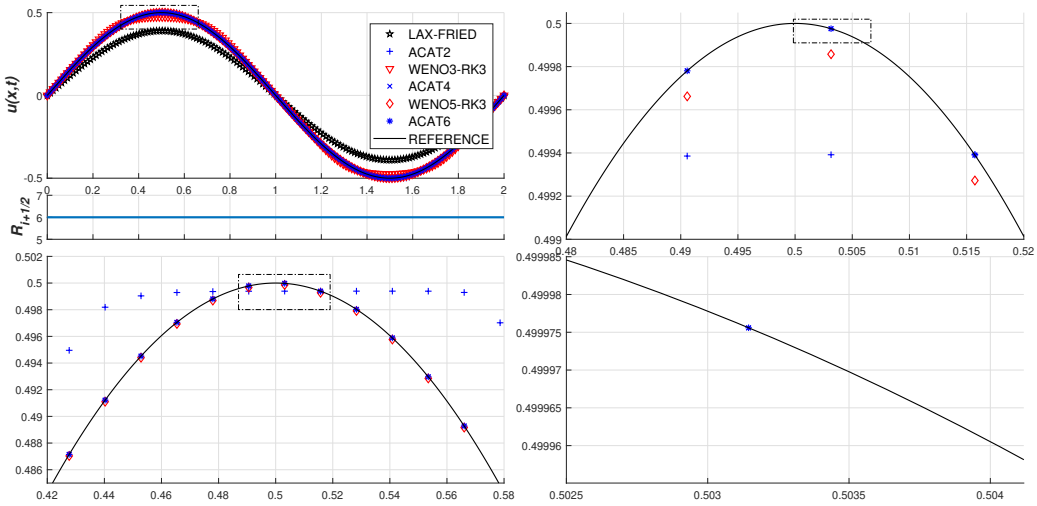

Figure 3.2.4: Transport equation with initial condition (3.2.24). Numerical solutions at $t = 40$: general view (*left-top*); local order of accuracy for ACAT6 (*sub-frame*);consecutive zooms close to the local maximum (*left-bottom*, *right-top* and *right-bottom*).

accuracy close to the critical points are related to the Flux-Limiter method and in particular how the flux limiter functions are computed; indeed, this method are not able to distinguish between critical or discontinuity points. This behaviour can also be observed for WENO-RK methods, although, in this case, this drawback can be overcome by using optimal weights in the WENO reconstructions: see [2, 3].

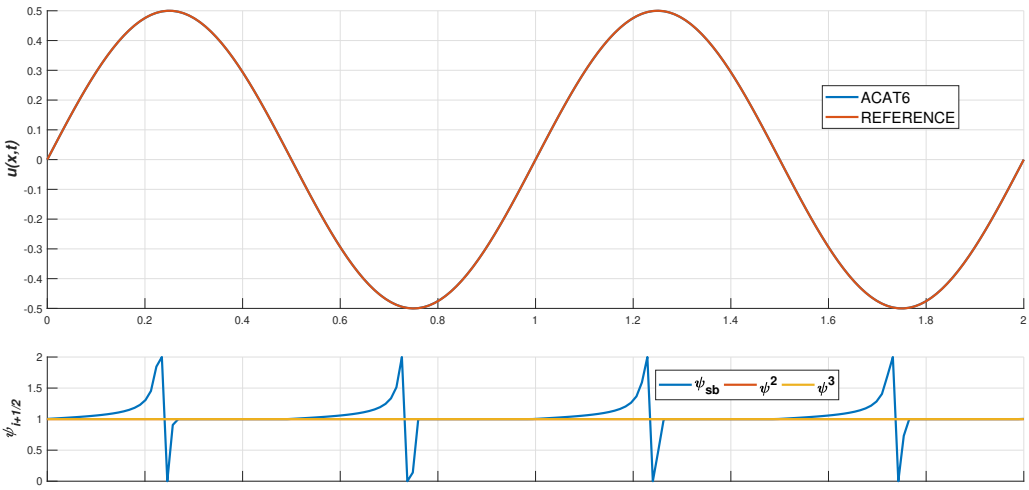

Figure 3.2.5: Transport equation with initial condition (3.2.25). Numerical solution obtained with ACAT6 at time $t = 3$ (*top*) and plot of the smoothness indicators ψ_{sb}, ψ^2 and ψ^3 (*bottom*).

Furthermore, the loss of accuracy of ACAT2 close to the critical points compared to

ACAT4 or 6 is due to the fact that, while the smoothness indicators $\psi^2_{i+\frac{1}{2}}$ and $\psi^3_{i+\frac{1}{2}}$ are always close to one, the Superbee flux limiter $\psi^1_{i+\frac{1}{2}} = \varphi^{\text{sb}}_{i+\frac{1}{2}}$ detects a discontinuity instead of a critical points and the first order method is then locally used. In order to emphasyze this behaviour, Figure 3.2.5 (top) shows the solution obtained with ACAT6 at time $t = 3$ for (3.2.23) with initial condition

$$u_0(x) = \frac{1}{2}\sin(2\pi x) \tag{3.2.25}$$

in the interval $[0, 2]$ using 160 mesh points, CFL= 0.9, and periodic boundary conditions. Figure 3.2.5 (*bottom*) exhibits the graph of the three smoothness indicators, the flux-limiter function ψ_{sb} and the second and fourth order smoothness indicators.

Finally, we notice that, for long time, the ACAT2 solution tents to be squared in the neighborhood of the critical points, see Figure 3.2.4. This behaviour is consequence, again, of the flux-limiter function $\varphi^1_{i+\frac{1}{2}}$. Indeed, as it shown on Figure 3.2.5, φ_{sb} is close to 0 in the critical points but there exist an interval $]x_c - \varepsilon, x_c[$, with $\varepsilon > 0$ and $x_c =$ critical point, where the flux-limiter function is bigger than 1.

Test 2: We consider again the transport equation (3.2.23) with a piecewise continuous initial condition

$$u_0(x) = \begin{cases} 1 & \text{if} \quad \frac{1}{2} \le x \le 1; \\ 0 & \text{if} \quad 0 \le x < \frac{1}{2} \quad \text{or} \quad \frac{3}{2} < x \le 2; \\ -1 & \text{if} \quad 1 < x \le \frac{3}{2}. \end{cases} \tag{3.2.26}$$

We solve numerically this problem in the spatial interval $[0, 2]$, using again 160 mesh points, CFL= 0.9, and periodic boundary conditions.

Figure 3.2.6 shows the solutions from ACAT2P, $P = 2, 4, 6$ and WENO$q-$RK3, $q = 3, 5$ at time $t = 2$ and $t = 20$. We can observe that ACAT2P produce less diffusive solutions than WENO$q-$RK3 in proximity of the shocks and this behaviour is emphasized when the methods are applied to a large time interval. Furthermore, the sub-frames show that larger is the time interval smaller is the $6^{th}-$order region.

Figures 3.2.7 and 3.2.8 show the errors vs CPU time plot with different initial conditions and different CFL-conditions. In practice, on Figure 3.2.7 can be seen that, in case of smooth initial condition, ACAT2P, $p = 1, 2$, WENOq-RK3, $q = 3, 5$, and Lax-Friedrichs methods are very similar in computational time vs errors sense but ACAT6 is faster and with lower

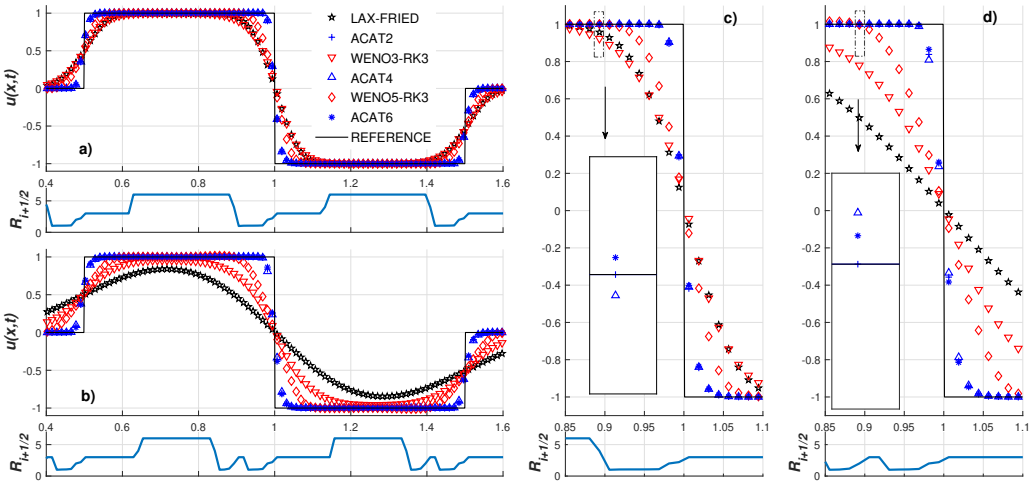

Figure 3.2.6: Transport equation with initial condition (3.2.26). Numerical solutions at $t = 2$ (**a**)) and at $t = 20$ (**b**)). Zooms of the numerical solutions close to the shock at time $t = 2$ (**c**)) and $t = 20$ (**d**)). Sub-frames: local order of accuracy for ACAT6.

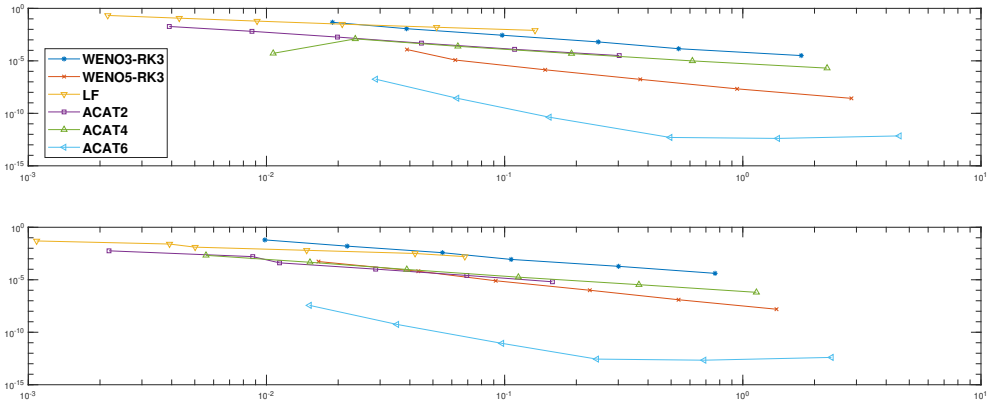

Figure 3.2.7: Error vs CPU time for the transport equation with smooth initial condition (3.2.24). Numerical solutions at $t = 4$ using CFL$= 0.5$ (*top*) and CFL$= 0.9$ (*bottom*).

error even for CFL $\in \{0.5, 0.9\}$. This behaviour is not maintained when they are applied to no-smooth initial condition because ACAT2P, for $p = 2, 3$, reduce to ACAT2 close to a discontinuity due to the adaptive procedure. Finally, Tables 3.3-3.6 display the errors in L^1-norm provided by the numerical solutions of WENO3-RK3, WENO5-RK3, LF, ACAT2, ACAT4 and ACAT6 methods supposing: $t = 4$; initial conditions smooth (3.2.25) and no-smooth (3.2.26); periodic boundary conditions; N $= \{50, 100, 200, 400, 800, 1600\}$ point meshes; and CFL $\in \{0.5, 0.9\}$. The reference solution is the exact one.

The following conclusions can be drawn:

- For smooth solutions: as expected, the errors decrease with the order of the methods.

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	4.6e-2	1.1e-4	2.1e-1	1.8e-2	5.2e-2	1.7e-7
100	1.1e-2	1.1e-5	1.1e-1	6.4e-3	1.2e-3	2.7e-9
200	2.7e-3	1.4e-6	6.0e-2	1.8e-3	2.4e-4	4.2e-11
400	6.3e-4	1.7e-7	3.1e-2	4.8e-4	4.9e-5	5.0e-13
800	1.4e-4	2.1e-8	1.5e-2	1.2e-4	6.8e-6	4.0e-13
1600	3.1e-5	2.6e-9	7.9e-3	3.1e-5	1.9e-7	1.2e-13

Table 3.3: Errors in L^1-norm for the transport equation (3.2.23) at time $t = 4$; smooth initial condition (3.2.24) and CFL= 0.5.

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	6.2e-2	5.4e-4	4.9e-2	5.7e-3	2.1e-3	3.6e-8
100	1.5e-2	6.4e-5	2.5e-2	1.5e-3	4.6e-4	5.0e-10
200	3.8e-3	8.0e-6	1.2e-2	4.0e-4	9.5e-5	5.5e-11
400	8.6e-4	1.0e-6	6.3e-3	1.0e-4	1.7e-5	2.7e-13
800	1.9e-4	1.2e-7	3.2e-3	2.5e-5	3.4e-6	2.1e-13
1600	4.0e-5	1.5e-8	1.6e-3	6.4e-6	4.3e-7	1.7e-13

Table 3.4: Errors in L^1-norm for the transport equation (3.2.23) at time $t = 4$; smooth initial condition (3.2.24) and CFL= 0.9.

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	4.7e-1	2.5e-1	8.3e-1	1.4e-1	1.4e-1	1.4e-1
100	2.6e-1	1.4e-1	6.3e-1	7.1e-2	7.1e-2	7.1e-2
200	1.5e-1	8.1e-2	4.5e-1	3.5e-2	3.5e-2	3.5e-2
400	9.5e-2	4.5e-2	3.2e-1	1.7e-2	1.7e-2	1.7e-2
800	5.7e-2	2.5e-2	2.2e-1	8.9e-3	8.9e-3	8.9e-3
1600	3.4e-2	1.4e-2	1.6e-1	4.4e-3	4.4e-3	4.4e-3

Table 3.5: Errors in L^1-norm for the transport equation (3.2.23) at time $t = 4$; smooth initial condition (3.2.26) and CFL= 0.5.

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	5.4e-1	3.1e-1	4.0e-1	1.2e-1	1.2e-1	1.2e-1
100	2.9e-1	2.8e-1	2.8e-1	6.6e-2	6.7e-2	6.8e-2
200	1.7e-1	2.4e-1	2.0e-1	3.5e-2	3.6e-2	3.6e-2
400	1.0e-1	1.7e-1	1.4e-1	1.8e-2	1.9e-2	1.9e-2
800	5.9e-2	1.3e-1	1.0e-1	9.3e-3	9.8e-3	9.5e-3
1600	3.4e-2	9.0e-2	7.2e-2	4.7e-3	5.0e-3	4.7e-3

Table 3.6: Errors in L^1-norm for the transport equation (3.2.23) at time $t = 4$; smooth initial condition (3.2.26) and CFL= 0.9.

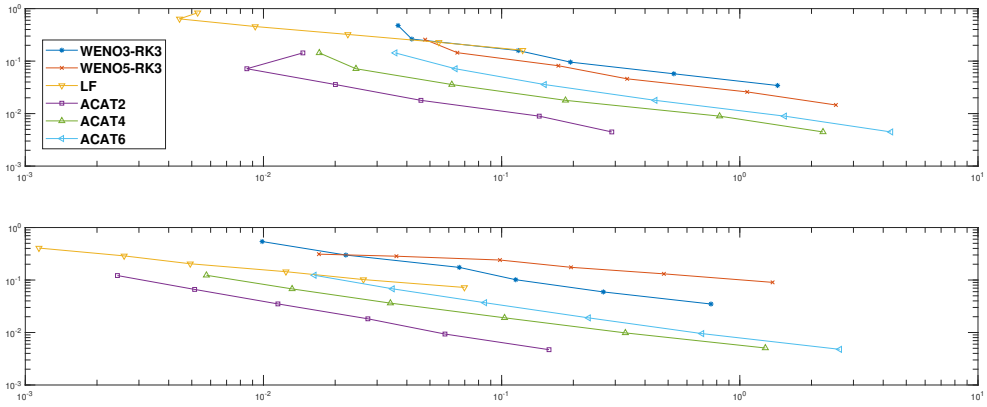

Figure 3.2.8: Error vs CPU time for the transport equation with no-smooth initial condition (3.2.26). Numerical solutions at $t = 4$ using CFL= 0.5 (*top*) and CFL= 0.9 (*bottom*).

Nevertheless, there is one exception: the second-order ACAT2 gives always lower errors than the third-order WENO3-RK3. The change of CFL from 0.5 to 0.9 does not significantly influence the behavior of the errors.

- For discontinuous solutions: ACAT methods give always lower errors than WENO-RK schemes. The errors corresponding to ACAT4 and ACAT6 are equal to those given by ACAT2 due to the fact since they both reduce to ACAT2 at the discontinuities due to the adaptive technique. WENO methods give bigger errors for CFL = 0.9 than for CFL = 0.5 due to the spurious oscillations appearing with the former value.

- The most efficient methods are ACAT6 for smooth solutions and ACAT2 for discontinuous ones.

Burgers equation

Let us consider the Burgers equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \tag{3.2.27}$$

with smooth initial condition (3.2.24). The problem is numerically solved in the interval $[0, 2]$ using a uniform mesh of 160 points, CFL= 0.9, and periodic boundary conditions.

Figures 3.2.9 and 3.2.10 show respectively the general view and the critical part zoom of the numerical solutions obtained with the different methods at times $t = \{0.25, 0.5, 1, 10\}$. On sub-frames, the local order of accuracy of ACAT6 is also displayed: as it can be seen, this method reduces to first order only at the shock once it has been generated.

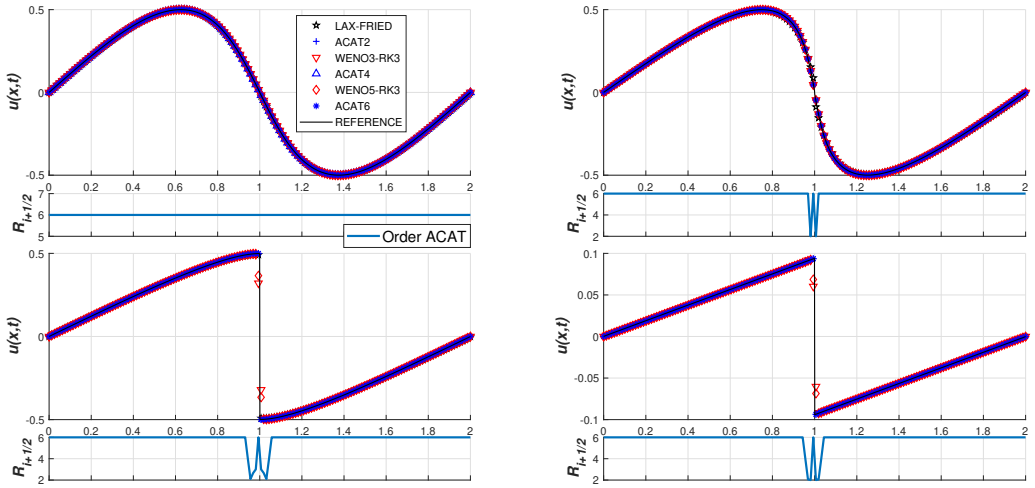

Figure 3.2.9: Burgers equation with smooth condition (3.2.24). Numerical solutions obtained at times $t = 0.25$ (*left-top*), $t = 0.5$ (*right-top*), $t = 1$ (*left-bottom*), and $t = 10$ (*right-bottom*). Sub-frames: local accuracy order for ACAT6.

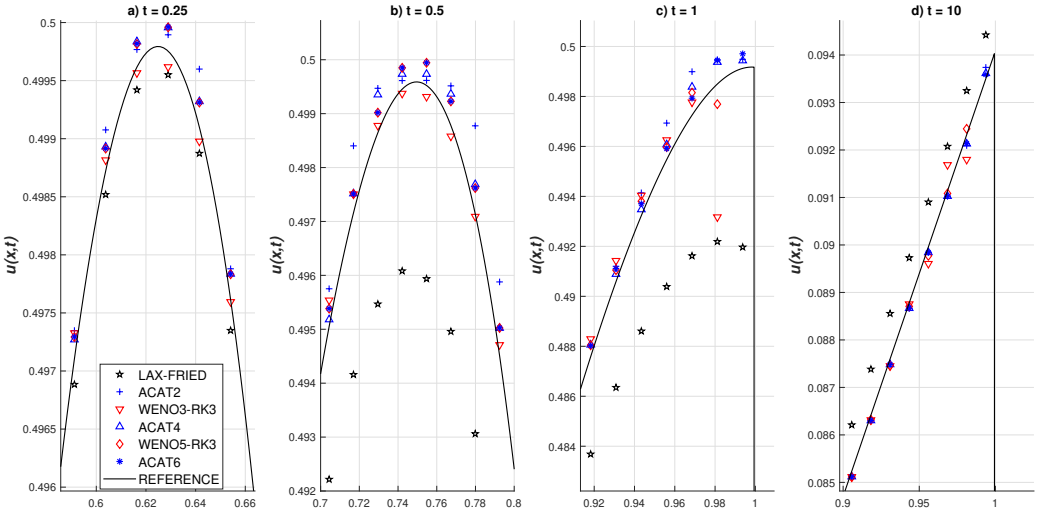

Figure 3.2.10: Burgers equation with smooth initial condition (3.2.24). Zoom of the numerical solutions obtained at times $t = 0.25$ **(a)**), $t = 0.5$ **(b)**), $t = 1$ **(c)**), and $t = 10$ **(d)**).

Figure 3.2.11 exhibits the errors vs CPU time plot with different CFL-conditions. We can observe that ACAT procedures are faster and with lower error than WENO-RK schemes, recording also that ACAT2P, $P = 2, 3$, reduce to ACAT2 in the region in which the discontinuity has been generated. Furthermore, Tables 3.7-3.8 show the errors in L^1-norm corresponding to the numerical solutions of WENO3-RK3, WENO5-RK3, LF, ACAT2, ACAT4 and ACAT6 methods supposing: $t = 4$; smooth initial condition (3.2.24); periodic boundary conditions; N $= \{50, 100, 200, 400, 800, 1600\}$ point meshes; and CFL $\in \{0.5, 0.9\}$. A refer-

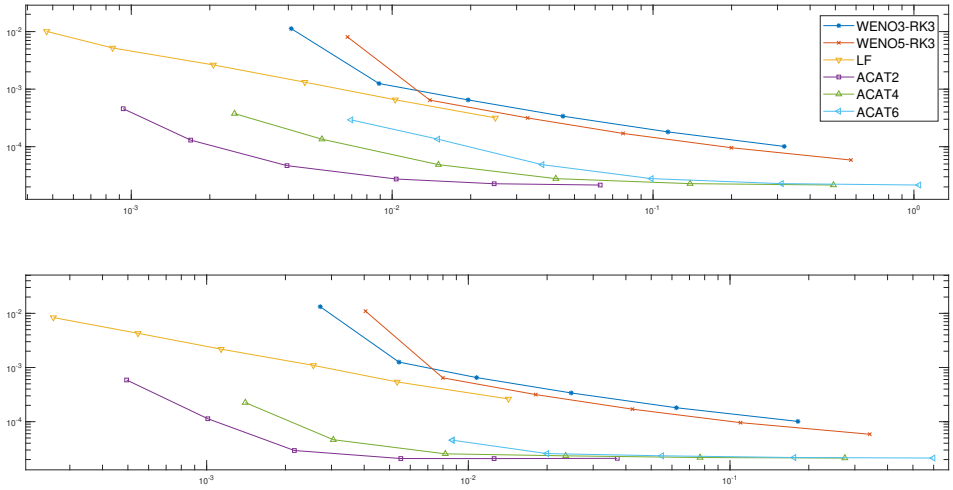

Figure 3.2.11: Error vs CPU time for the burgers equation with smooth initial condition (3.2.24). Numerical solutions at $t = 2$ adopting, respectively, CFL= 0.5 (*top*) and CFL= 0.9 (*bottom*).

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	1.1e-2	8.0e-3	1.0e-2	4.5e-4	3.7e-4	2.9e-4
100	1.2e-3	6.4e-4	5.1e-3	1.3e-4	1.3e-4	1.3e-4
200	6.4e-4	3.1e-4	2.6e-3	4.6e-5	4.5e-5	4.5e-5
400	3.3e-4	1.6e-4	1.3e-3	2.7e-5	2.7e-5	2.7e-5
800	1.8e-4	9.5e-5	6.5e-4	2.2e-5	2.2e-5	2.2e-5
1600	1.0e-4	8.6e-6	3.1e-4	2.1e-5	2.0e-5	2.0e-5

Table 3.7: Errors in L^1-norm for the Burgers equation with smooth initial condition (3.2.24). Numerical solutions at $t = 4$ using CFL= 0.5.

ence solution has been computed with 3200 mesh points, so that the numerical solution is compared with a reference solution on grid points with the same abscissa. The reference solution plotted in Figures 3.2.9 and 3.2.10 has been computed with 1400 grid points. The conclusions are similar to those obtained for **Test 2**.

1D Euler equations

Let us now skip to systems of conservation laws. In particular, we will focus on the 1D Euler equations for gas dynamics

$$U_t + f(U)_x = 0, \tag{3.2.28}$$

N	W3RK3	W5RK3	LF	ACAT2	ACAT4	ACAT6
50	1.3e-2	1.0e-2	8.3e-3	5.8e-4	2.2e-4	2.0e-5
100	1.2e-3	6.4e-4	4.2e-3	1.1e-4	4.6e-5	4.5e-5
200	6.5e-4	3.1e-4	2.1e-3	2.9e-5	2.5e-5	2.5e-5
400	3.3e-4	1.7e-4	1.0e-3	2.0e-5	2.3e-5	2.3e-5
800	1.8e-4	9.5e-5	5.4e-4	2.0e-5	2.1e-5	2.1e-5
1600	1.0e-4	5.8e-5	2.6e-4	2.1e-5	2.0e-5	2.0e-5

Table 3.8: Errors in L^1-norm for the Burgers equation with smooth initial condition (3.2.24). Numerical solutions at $t = 4$ using CFL= 0.9.

with

$$U = \begin{bmatrix} \rho \\ \rho v \\ E \end{bmatrix}, \quad f(U) = \begin{bmatrix} \rho v \\ p + \rho v^2 \\ v(E + p) \end{bmatrix}, \tag{3.2.29}$$

where ρ is the density measured in kg/m^3; v, the velocity in m/s; E the total energy per unit volume in $Kg/(ms^2)$; and p is the pressure in Pascal Pa. We assume an ideal gas with the equation of state

$$p(\rho, e) = (\gamma - 1)\rho e,$$

being γ the ratio of specific heat capacities of the gas taken as 1.4 and e is the internal energy per unit mass related to E by:

$$E = \rho(e + 0.5v^2).$$

We consider four Riemann problems for (3.2.28): the Sod problem [114]; the Einfeldt problem [36]; the right blast wave Woodward and Colella problem [128]; and the Shu-Osher problem [112]. In the first three cases: the initial discontinuity is placed at $x = 0.5$ and the equations are numerically solved at the spatial interval $[0, 1]$. For the Shu-Osher problem the initial discontinuity is set at $x = -4$ and the equations are solved numerically on the interval $[-5, 5]$. In all cases the exact solution is provided by the HE-E1RPEXACT solver introduced in [120]; the CFL parameter is set to 0.8 and outflow-inflow boundary conditions are considered.

The Sod problem:

$$(\rho, v, p) = \begin{cases} (1, 0, 1) & \text{if } x < 1/2, \\ (0.125, 0, 0.1) & \text{if } x > 1/2. \end{cases} \tag{3.2.30}$$

The solution involves a rarefaction wave, a contact discontinuity and a shock. We compare

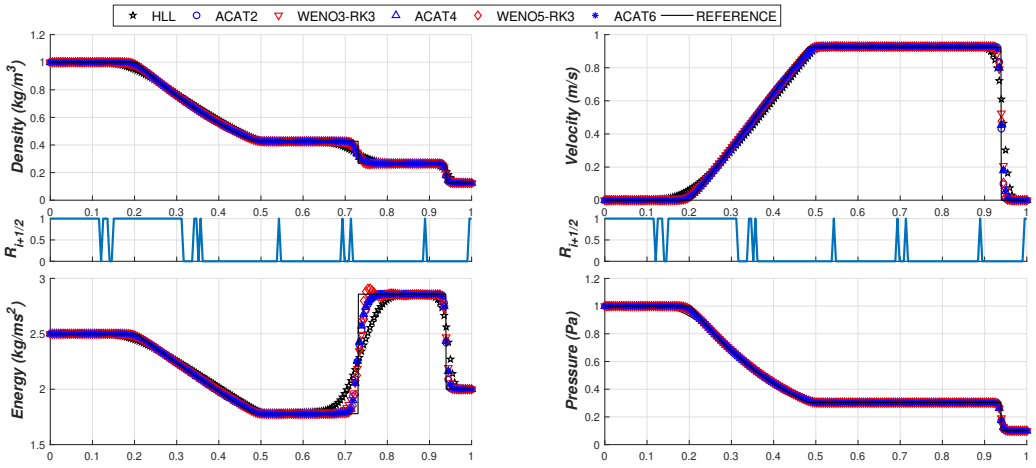

Figure 3.2.12: 1D Euler equations: the Sod problem. Numerical solutions at $t = 0.25$ using CFL= 0.8 and 200 points: density (*left-top*), velocity (*right-top*), internal energy (*left-bottom*), pressure (*right-bottom*). Sub-frames: local order of accuracy for ACAT6.

the numerical solutions with the exact one: see [120]. Figure 3.2.12 shows the solutions provided by ACAT2P, $P = 1, 2, 3$, WENO$q-$RK3, $q = 3, 5$, and HLL for density, velocity, internal energy and pressure p, using 200 mesh points. The local accuracy of ACAT6 is also shown. Zooms of the behaviour of the numerical densities can be observed in Figure 3.2.13. As it can be seen in zooms a and b, WENO5-RK3 gives sharper but more oscillatory solutions than ACAT methods. Moreover, increasing the accuracy order for ACAT methods we obtain sharper results. Similar conclusions for the internal energy can be drawn: see Figure 3.2.14.

	ρ	ρv	E	ρ	ρv	E	ρ	ρv	E
N	W3RK3			W5RK3			LF		
50	1.6e-2	1.4e-2	3.7e-2	1.2e-2	1.0e-2	2.7e-2	3.0e-2	2.8e-2	7.0e-2
100	8.5e-3	8.2e-3	1.9e-2	6.5e-3	6.5e-3	1.5e-2	1.9e-2	1.8e-2	4.1e-2
200	4.3e-3	4.2e-3	9.4e-3	3.2e-3	3.2e-3	7.5e-3	1.2e-2	1.1e-2	2.0e-2
400	2.1e-3	2.0e-3	4.6e-3	1.5e-3	1.5e-3	3.7e-3	7.3e-3	6.7e-3	1.2e-2
800	9.5e-4	9.6e-4	2.2e-3	7.6e-4	7.7e-4	1.8e-3	4.4e-3	4.0e-3	8.2e-3
1600	4.7e-4	4.2e-4	9.4e-4	4.7e-4	4.4e-4	9.1e-4	2.5e-3	2.3e-3	4.5e-3

Table 3.9: 1D Euler equations: Sod problem. Errors in L^1-norm for ρ, ρv and E computed with WENOq-RK3, $q = 3, 5$, and Lax-Friedrichs at time $t = 0.25$ using CFL= 0.5.

Tables 3.9-3.10 exhibit the error in L^1-norm for density ρ, momentum ρv and energy E computed with ACAT2P, $P = 1, 2, 3$, WENO$q-$RK3, $q = 3, 5$, and HLL first order method

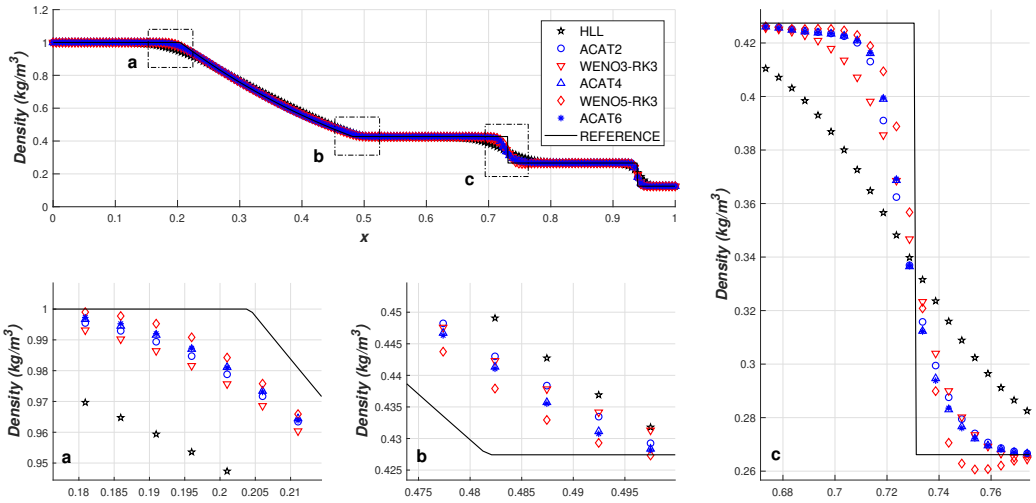

Figure 3.2.13: 1D Euler equations: the Sod problem. Numerical densities at $t = 0.25$ using CFL= 0.8 and 200 points: general view and zooms close to the points *a*,*b*, *c* and *d*.

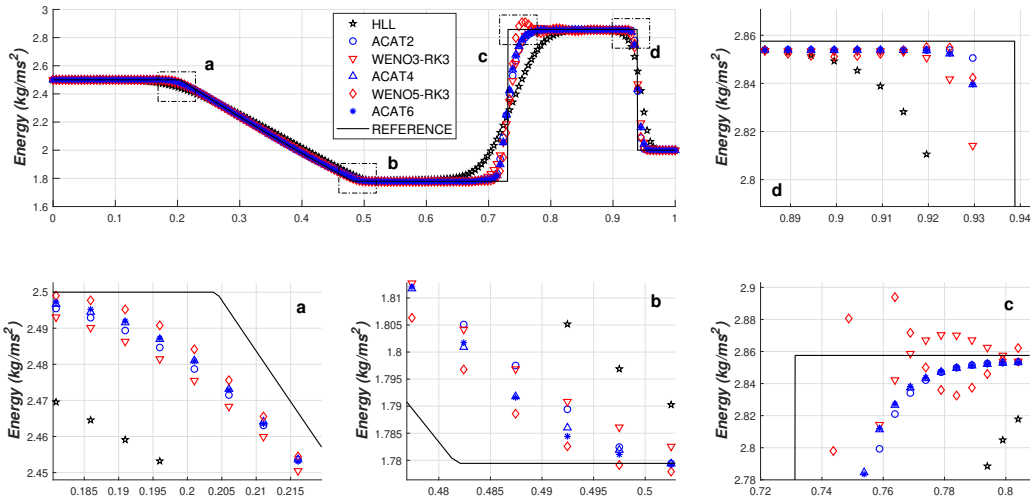

Figure 3.2.14: 1D Euler equations: the Sod problem. Numerical internal energies at $t = 0.25$ using CFL= 0.8 and 200 points: general view and zooms close to the points *a*,*b*, *c* and *d*.

at time $t = 0.25$ adopting CFL= 0.5. What we can seen is that ACAT methods produce similar error as WENO procedure that seems to be in contrast with density and Energy plots, but in this table the CFL is set to 0.5 than the WENO reconstructions do not produce spurious oscillations.

Figure 3.2.15 shows the errors vs CPU times and errors in L^1-norm respectively corresponding to the numerical solutions of WENO3-RK3, WENO5-RK3, LF, ACAT2, ACAT4 and ACAT6 methods for the Sod problem using: $t = 0.25$; N $= \{50, 100, 200, 400, 800, 1600\}$

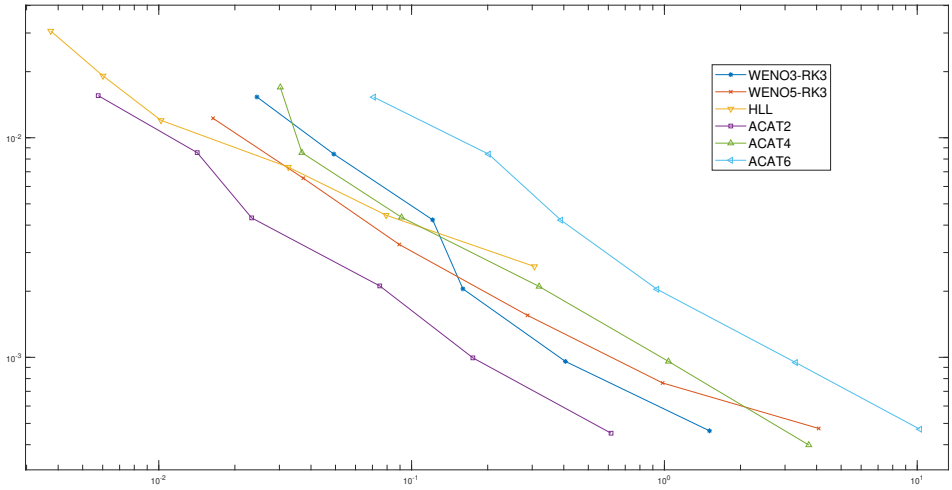

Figure 3.2.15: Error vs CPU time for the Sod problem. Numerical solutions at $t = 0.25$ using CFL= 0.5.

	ρ	ρv	E	ρ	ρv	E	ρ	ρv	E
N	ACAT2			ACAT4			ACAT6		
50	1.5e-2	1.3e-2	3.2e-2	1.5e-2	1.3e-2	3.2e-2	1.5e-2	1.3e-2	3.2e-2
100	8.5e-3	7.9e-3	1.7e-2	8.4e-3	7.9e-3	1.7e-2	8.4e-3	7.9e-3	1.7e-2
200	4.3e-3	4.0e-3	8.8e-3	4.2e-3	3.9e-3	8.8e-3	4.2e-3	3.9e-3	8.8e-3
400	2.1e-3	1.9e-3	4.4e-3	2.0e-3	1.9e-3	4.4e-3	2.0e-3	1.9e-3	4.4e-3
800	9.9e-4	9.5e-4	2.2e-3	9.5e-4	9.2e-4	2.2e-3	9.4e-4	9.1e-4	2.2e-3
1600	4.5e-4	4.0e-4	9.4e-4	4.6e-4	4.1e-4	9.3e-4	3.9e-4	3.8e-4	9.3e-4

Table 3.10: 1D Euler equations: Sod problem. Errors in L^1-norm for ρ, ρv and E computed with ACAT2P, $p = 1, 2, 3$, at time $t = 0.25$ using CFL= 0.5.

point meshes; and CFL = 0.5. The reference solution is the exact one provided by the HE-E1RPEXACT algorithm. The following conclusions can be drawn:

- The errors given by all the methods are comparable under CFL= 0.5.
- ACAT2 is the most efficient method, followed by WENO3-RK3; the efficiencies of ACAT4 and WENO5-RK3 are comparable; ACAT6 is the least efficient method in this case. Please note first of all that a more restricted CFL-condition is adopted; sencondly, due to the adaptivity property, the ACAT2P reduce to ACAT2 close a discontinuity; finally, a non-optimized Matlab implementation of the methods has been used to compute the numerical solutions. ACAT methods are highly parallelisable and do not need the storage of intermediate temporal stages, so that an optimized parallel implementation can lead to very different conclusions.

The 123 Einfeldt problem:

The solution of this problem involves two strong rarefaction waves and an intermediate state that is close to vacuum, what makes this problem a hard test for numerical methods.

$$(\rho, v, p) = \begin{cases} (1.0, -2.0, 0.4) & \text{if } x < 1/2, \\ (1.0, 2.0, 0.4) & \text{if } x > 1/2. \end{cases} \tag{3.2.31}$$

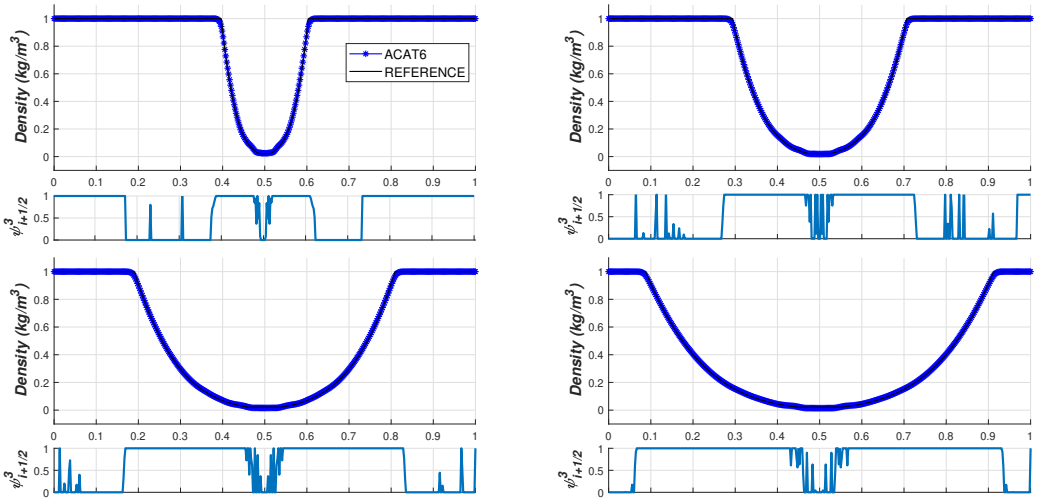

Figure 3.2.16: 1D Euler equations: the 123 Einfeldt problem. Numerical solutions at $t_s = 0.15$ using CFL= 0.8 and 200 points. Density obtained with ACAT6 and graph of the smoothness indicator ψ^3 for $t = t_s/4$ (*left-top*); $t_s/2$ (*right-top*); $3t_s/4$ (*left-bottom*); t_s (*right-bottom*).

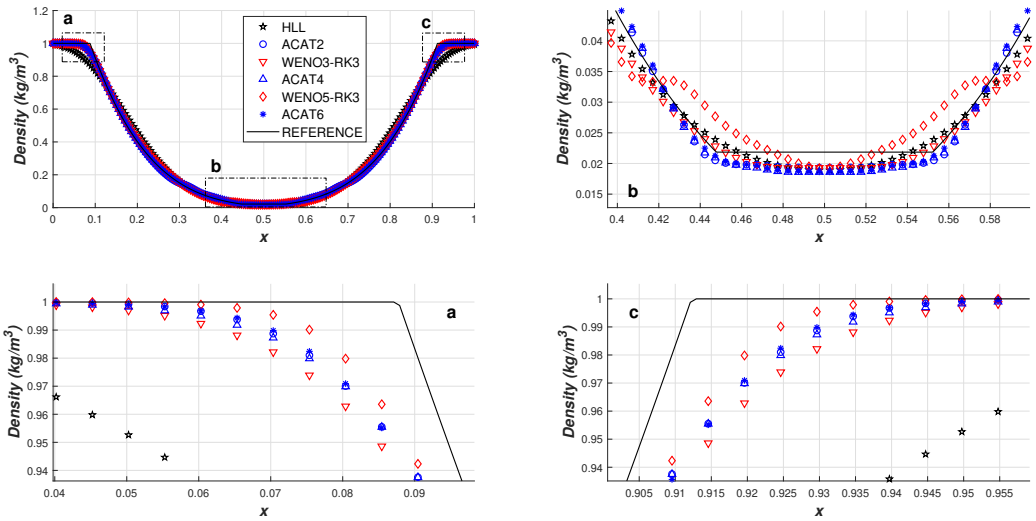

Figure 3.2.17: 1D Euler equations: the 123 Einfeldt problem. Numerical solutions at $t = 0.15$ using CFL= 0.8 and 200 points: general view (*left-top*). Zooms close to the points **a** (*left-bottom*), **b**(*right-top*), and **c** (*right-bottom*).

ACAT methods give stable solutions under CFL$\leq$ 1 condition: Figure 3.2.16 shows the

time evolution of the numerical results obtained with ACAT6. The smoothness indicators $\psi^3_{i+\frac{1}{2}}$ is also depicted: it can be seen how the discontinuities of the first order derivatives are correctly captured. It can be also observed that, while at the rarefaction waves order 6 is selected, lower accuracy is used at the constant regions close to the boundaries: this order reduction is due to the numerical oscillations produced by the 6th order method. A comparison of ACAT2P, $P = 1, 2, 3$, with different methods, WENOq−RK3, $q = 3, 5$, and HLL first order, at time $t = 0.15$ is shown in Figure 3.2.17 using 200 mesh points, where ACAT methods provide similar stable solutions. Although WENO solutions are stable, the third-order one is diffusive and the fifth-order one is oscillatory.

The right blast wave problem of Woodward & Colella:

The solution of this problems involves a rarefaction waves and two strong shock that make this an hard test for the numerical schemes.

$$(\rho, v, p) = \begin{cases} (1.0, 0.0, 1000) & \text{if } x < 1/2, \\ (1.0, 0.0, 0.01) & \text{if } x > 1/2. \end{cases} \tag{3.2.32}$$

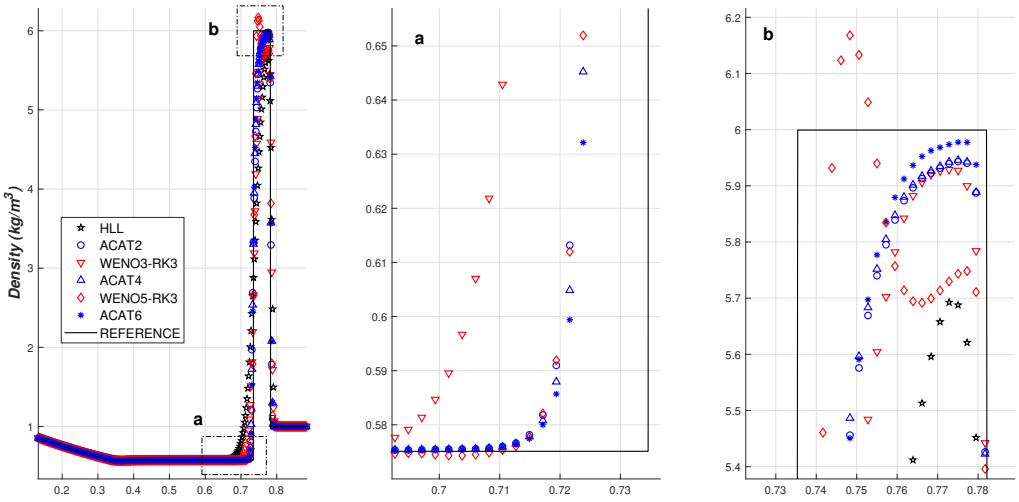

Figure 3.2.18: 1D Euler equations: right blast wave of the Woodward & Colella problem. Numerical solutions at time $t = 0.012$ using CFL= 0.8 and 450 points (*left*). Zooms close to the shocks (*center and right*).

For this tests we use 450 mesh points. The solution involves two strong shocks. Figure 3.2.18 shows the numerical densities obtained at time $t = 0.012$ with ACAT2P, $P = 1, 2, 3$, WENOq−RK3, $q = 3, 5$, and HLL schemes. It can be observed that WENO methods produce oscillating solutions, while ACAT methods give stable solutions whose accuracy

increase with the order. In particular, this behavior is emphasized on the two zooms close to the shocks.

The Shu-Osher problem: The solution of this problem concerns a strong shock with a wavelike initial condition that increases the difficulty of the problem numerically.

$$(\rho, v, p) = \begin{cases} (27/7, 2.629369, 10 + 1/3) & \text{if } x < -4, \\ (1 + \frac{1}{5}\sin(5\pi x), 0.0, 1) & \text{if } x > -4. \end{cases} \tag{3.2.33}$$

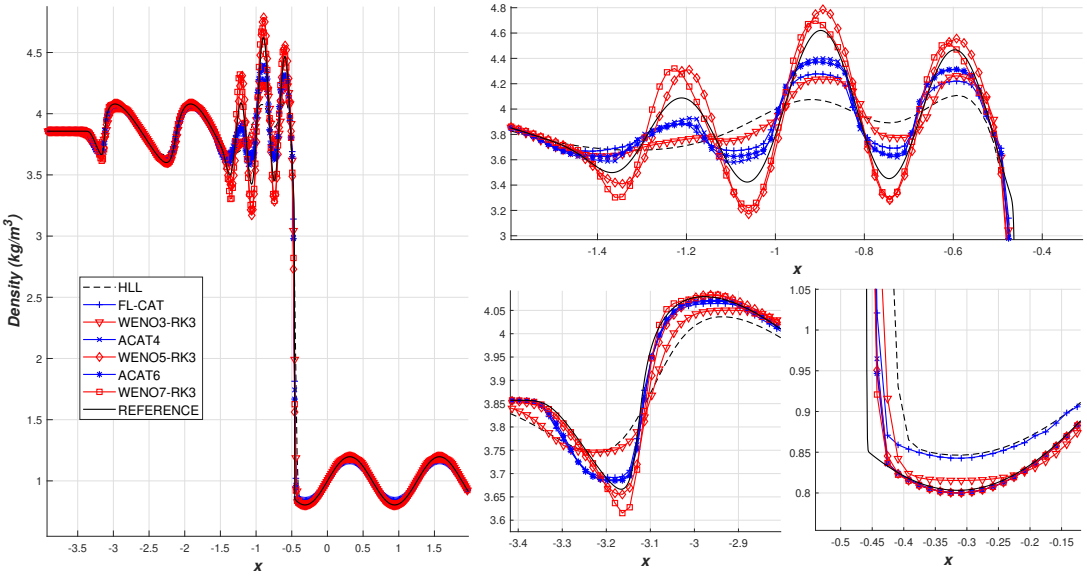

Figure 3.2.19: 1D Euler equations: Shu-Osher problem. Numerical solutions at time $t = 1$ using CFL= 0.8 and 450 points (*left*). Zooms close to the shock and wavelike parts (*center and right*).

Figure 3.2.19 shows the solution provided by ACAT2P, $P = 1, 2, 3$, WENOq−RK3, $q = 3, 5, 7$ and HLL methods for density using 450 mesh points. We observe that first order HLL is very diffusive in the wavelike region; ACAT procedures present some diffusion close the undulating part and it seems that WENO5 − 7 gives better result but they also overshoot the exact solution.

3.3 2D Adaptive Compact Approximate Taylor Method

In this section we focus on the extension of ACAT methods to non-linear two-dimensional systems of hyperbolic conservation laws

$$U_t + f(U)_x + g(U)_y = 0. \tag{3.3.1}$$

The following multi-index notation will be used:

$$\mathbf{i} = (i_1, i_2) \in \mathbb{Z} \times \mathbb{Z},$$

and

$$\mathbf{0} = (0,0), \quad \mathbf{1} = (1,1), \quad \frac{\mathbf{1}}{\mathbf{2}} = (\frac{1}{2}, \frac{1}{2}), \quad \mathbf{e}_1 = (1,0), \quad \mathbf{e}_2 = (0,1).$$

We consider Cartesian meshes with nodes

$$\mathbf{x_i} = (i_1 \Delta x, i_2 \Delta y).$$

Using this notation, the general form of the CAT2p method will be as follows:

$$U_{\mathbf{i}}^{n+1} = U_{\mathbf{i}}^{n} + \frac{\Delta t}{\Delta x}\left[F^p_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}\right] + \frac{\Delta t}{\Delta y}\left[G^p_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}\right], \tag{3.3.2}$$

where the numerical fluxes $F^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ will be computed using the values of the numerical solution $U_{\mathbf{i}}^n$ in the p^2-point stencil centered at $\mathbf{x}_{\mathbf{i}+\frac{1}{2}} = ((i_1 + \frac{1}{2})\Delta x, (i_2 + \frac{1}{2})\Delta y)$

$$S^p_{\mathbf{i}+\frac{1}{2}} = \{\mathbf{x}_{\mathbf{i}+\mathbf{j}}, \quad \mathbf{j} \in \mathcal{I}_p\},$$

where

$$\mathcal{I}_p = \{\mathbf{j} = (j_1, j_2) \in \mathbb{Z} \times \mathbb{Z}, \quad -p+1 \le j_k \le p, \quad k = 1,2\}.$$

See Figure 3.3.1 for an idea of the 2D meshes for $p = 2$.

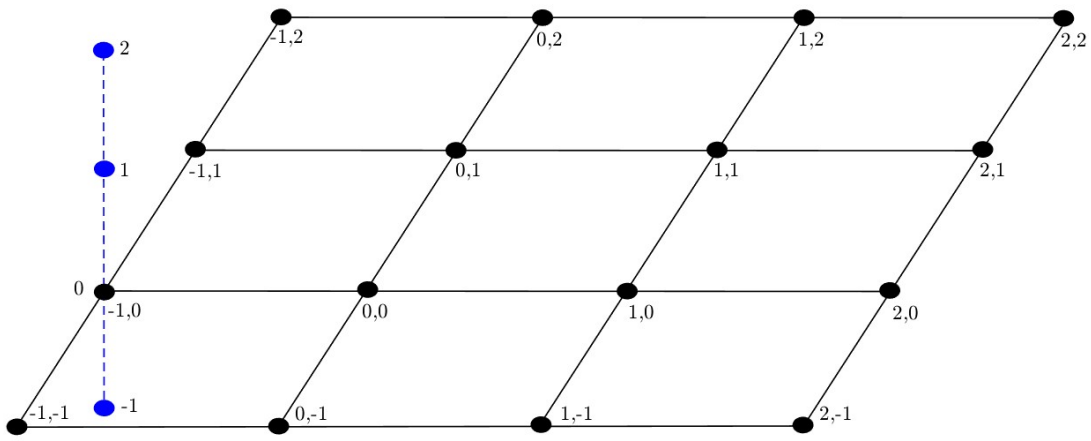

Figure 3.3.1: Stencil $S^2_{\mathbf{i}+\frac{1}{2}}$ centered in $\mathbf{x}_{\frac{1}{2}} = \frac{1}{2}(\Delta x, \Delta y)$

3.3.1 2D CAT2

In order to show the extension of CAT2P procedure let us start with the expression of the CAT2. In particular taking in mind Figure 3.3.1, the numerical fluxes are constructed as follows:

$$F^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \frac{1}{4}\left(f^{1,n+1}_{\mathbf{i},\mathbf{0}} + f^{1,n+1}_{\mathbf{i},\mathbf{e}_1} + f^n_{\mathbf{i}} + f^n_{\mathbf{i}+\mathbf{e}_1}\right), \tag{3.3.3}$$

$$G^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \frac{1}{4}\left(g^{1,n+1}_{\mathbf{i},\mathbf{0}} + g^{1,n+1}_{\mathbf{i},\mathbf{e}_2} + g^n_{\mathbf{i}} + g^n_{\mathbf{i}+\mathbf{e}_2}\right), \tag{3.3.4}$$

where

$$f^{1,n+1}_{\mathbf{i},\mathbf{j}} = f\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i},\mathbf{j}}\right),$$
$$g^{1,n+1}_{\mathbf{i},\mathbf{j}} = g\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i},\mathbf{j}}\right),$$

for $\mathbf{j} = \mathbf{0}, \mathbf{e}_1$ in the x direction, and $\mathbf{j} = \mathbf{0}, \mathbf{e}_2$ in the y direction.

Remark 3.3.1 *Despite what happen for the 1D reconstruction, the first time derivative of U, $U^{(1)}_{\mathbf{i},\mathbf{j}}$, does not coincide in the 2D-grid points. Indeed, observe that $U^{(1)}_{\mathbf{i},\mathbf{0}} \neq U^{(1)}_{\mathbf{i},\mathbf{e}_1}$ and $U^{(1)}_{\mathbf{i},\mathbf{0}} \neq U^{(1)}_{\mathbf{i},\mathbf{e}_2}$.*

Note that, in the 1D case, $U^{(1)}_{i,0} = U^{(1)}_{i,1}$ as in step 2.

Hence, the first time derivatives $U^{(1)}_{\mathbf{i},\mathbf{j}}$ are so defined:

$$U^{(1)}_{\mathbf{i},\mathbf{0}} = -\frac{1}{\Delta x}\left(f^n_{\mathbf{i}+\mathbf{e}_1} - f^n_{\mathbf{i}}\right) - \frac{1}{\Delta y}\left(g^n_{\mathbf{i}+\mathbf{e}_2} - g^n_{\mathbf{i}}\right),$$
$$U^{(1)}_{\mathbf{i},\mathbf{e}_1} = -\frac{1}{\Delta x}\left(f^n_{\mathbf{i}+\mathbf{e}_1} - f^n_{\mathbf{i}}\right) - \frac{1}{\Delta y}\left(g^n_{\mathbf{i}+\mathbf{1}} - g^n_{\mathbf{i}+\mathbf{e}_1}\right),$$
$$U^{(1)}_{\mathbf{i},\mathbf{e}_2} = -\frac{1}{\Delta x}\left(f^n_{\mathbf{i}+\mathbf{1}} - f^n_{\mathbf{i}+\mathbf{e}_2}\right) - \frac{1}{\Delta y}\left(g^n_{\mathbf{i}+\mathbf{e}_2} - g^n_{\mathbf{i}}\right),$$

where

$$f^n_{\mathbf{i}+\mathbf{j}} = f(U^n_{\mathbf{i}+\mathbf{j}}), \quad g^n_{\mathbf{i}+\mathbf{j}} = g(U^n_{\mathbf{i}+\mathbf{j}}), \quad \forall \mathbf{j}.$$

Finally, the 2D CAT2 method is so get:

$$U^{n+1}_{\mathbf{i}} = U^n_{\mathbf{i}} + \frac{\Delta t}{\Delta x}\left[F^1_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}\right] + \frac{\Delta t}{\Delta y}\left[G^1_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}\right], \tag{3.3.5}$$

3.3.2 2D CAT2P

The high order CAT2p iterative procedure are computed as following:

1. Define
$$f^{(0)}_{\mathbf{i},\mathbf{j}} = f^n_{\mathbf{i}+\mathbf{j}}, \quad g^{(0)}_{\mathbf{i},\mathbf{j}} = g^n_{\mathbf{i}+\mathbf{j}}, \quad \mathbf{j} \in \mathcal{I}_p.$$

2. For $k = 2, \ldots, 2p$:

   (a) Compute
$$U^{(k-1)}_{\mathbf{i},\mathbf{j}} = -A^{1,j_1}_p(f^{(k-2)}_{\mathbf{i},(*,j_2)}, \Delta x) - A^{1,j_2}_p(g^{(k-2)}_{\mathbf{i},(j_1,*)}, \Delta y), \quad \mathbf{j} \in \mathcal{I}_p.$$

   (b) Compute
$$f^{k-1,n+r}_{\mathbf{i},\mathbf{j}} = f\left(U^n_{\mathbf{i}+\mathbf{j}} + \sum_{l=1}^{k-1} \frac{(r\Delta t)^l}{l!} U^{(l)}_{\mathbf{i},\mathbf{j}}\right), \quad \mathbf{j} \in \mathcal{I}_p, \ \ r = -p+1, \ldots, p.$$

   (c) Compute
$$f^{(k-1)}_{\mathbf{i},\mathbf{j}} = A^{k-1,0}_p(f^{k-1,*}_{\mathbf{i},\mathbf{j}}, \Delta t), \quad \mathbf{j} \in \mathcal{I}_p.$$

3. Compute
$$F^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \sum_{k=1}^{2p} \frac{\Delta t^{k-1}}{k!} A^{0,1/2}_p(\tilde{f}^{(k-1)}_{\mathbf{i},(*,0)}, \Delta x), \tag{3.3.6}$$
$$G^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \sum_{k=1}^{2p} \frac{\Delta t^{k-1}}{k!} A^{0,1/2}_p(\tilde{g}^{(k-1)}_{\mathbf{i},(0,*)}, \Delta y). \tag{3.3.7}$$

The notation used for the approximation of the spacial partial derivatives is the following:

$$A^{k,q}_p(f_{\mathbf{i},(*,j_2)}, \Delta x) = \frac{1}{\Delta x^k} \sum_{l=-p+1}^{p} \gamma^{k,q}_{p,l} f_{\mathbf{i},(l,j_2)}$$
$$A^{k,q}_p(g_{\mathbf{i},(j_1,*)}, \Delta y) = \frac{1}{\Delta y^k} \sum_{l=-p+1}^{p} \gamma^{k,q}_{p,l} g_{\mathbf{i},(j_1,l)}$$

Remark 3.3.2 *In the last step of the algorithm above the set $\mathcal{I}_p$ can be replaced by its $(2p-1)$-point subset*

$$\mathcal{I}^0_p = \{\mathbf{j} = (j_1, j_2) \ \text{ such that } \ j_1 = 0 \ \text{ or } \ j_2 = 0\}$$

since only the corresponding values of $\tilde{f}_{\mathbf{i},\mathbf{j}}^{(k-1)}$ *are used to compute the numerical fluxes* (3.3.6) *and* (3.3.7).

3.3.3 2D ACAT2P

Once the numerical flux of the CAT2p method has been introduced, the numerical flux ACAT2 is extended to the two-dimensional problems through the flux-limiter scheme as follows:

$$F^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \varphi_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} F^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + (1-\varphi 1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}) F^{lo}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}, \tag{3.3.8}$$

$$G^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \varphi_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} G^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + (1-\varphi_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}) G^{lo}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}, \tag{3.3.9}$$

where, $F^{lo}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$ and $G^{lo}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ are some robust first order methods; $\varphi_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$ and $\varphi_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ are the flux limiters computed dimension by dimension.

Finally, the expression of the ACAT2P method for two-dimensional problems is

$$U^{n+1}_{\mathbf{i}} = U^n_{\mathbf{i}} + \frac{\Delta t}{\Delta x}\left(F^{A_1}_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^{A_1}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}\right) + \frac{\Delta t}{\Delta y}\left(G^{A_2}_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^{A_2}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}\right), \tag{3.3.10}$$

where the numerical fluxes are so defined:
firstly, let us consider the sets

$$\mathcal{A}_1 = \{p \in \{2,\dots,P\} \text{ such that } \psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} \approx 1\}, \tag{3.3.11}$$

$$\mathcal{A}_2 = \{p \in \{2,\dots,P\} \text{ such that } \psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} \approx 1\}, \tag{3.3.12}$$

where $\psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $\psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ are the smoothness indicators introduced in Section 3.2.2 computed dimension by dimension. Then define:

$$F^{A_1}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \begin{cases} F^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{if } \mathcal{A}_1 = \emptyset; \\ F^{p_1}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{where } p_1 = \max(\mathcal{A}_1) \text{ otherwise;} \end{cases} \tag{3.3.13}$$

$$G^{A_2}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \begin{cases} G^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{if } \mathcal{A}_2 = \emptyset; \\ G^{p_2}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{where } p_2 = \max(\mathcal{A}_2) \text{ otherwise.} \end{cases} \tag{3.3.14}$$

Observe that, since the smoothness indicators are computed dimension by dimension, a

rectangular stencil

$$S^{p_1,p_2}_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}} = \{\mathbf{x_{i,j}}, \quad i_1 - p_1 + 1 \leq j_1 \leq i_1 + p_1, \quad i_2 - p_2 + 1 \leq j_2 \leq i_2 + p_2\},$$

is used in practice to compute the numerical fluxes $F^{p_1}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^{p_2}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$. The extension of CAT methods to such rectangular stencils is straightforward.

3.3.4 Numerical experiments

In this section we focus on the behaviour of the ACAT2P procedure applied to 2D problem: scalar transport equation and the 2D Euler equation for gas dynamic. As flux limiter function, for 2D ACAT2 method, the minmod flux limiter [120] is used direction by direction; and the smoothness indicators (3.2.9) are considered direction by direction. ACAT2P is again compared with WENO($2p+1$) finite difference methods based on the Lax-Friedrichs splitting techniques (see [110]) combined with SSPRK3 [48] for the time reconstruction.

2D Transport equation

Let us consider the 2D transport equation

$$u_t + au_x + bu_y = 0, \tag{3.3.15}$$

with initial condition

$$u = \begin{cases} 1 & \text{if } x + y \leq 1/4, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3.16}$$

We solve (3.3.15) on the spatial domain $[0,2]\times[0,2]$, using: $a, b = 1$, 100×100-point grid, CFL= 0.5, free boundary conditions and $t = 1$. Figure 3.3.2 shows a 1D cut over the first diagonal $y = x$ of the numerical solutions obtained with ACAT2, ACAT4, WENO3-RK3 and WENO5-RK3 at time $t = 1$.

2D Euler equations

Let us consider the two-dimensional Euler equations for gas dynamics

$$U_t + f(U)_x + g(U)_y = 0, \tag{3.3.17}$$

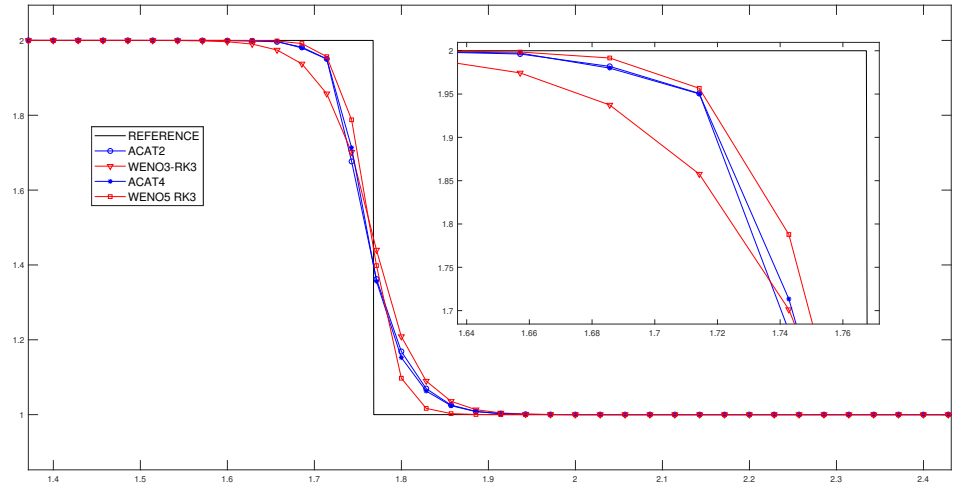

Figure 3.3.2: 2D Transport equation. Solutions obtained with ACAT2, ACAT4, WENO3-RK3 and WENO5-RK3 at time $t = 1$: cut with a vertical plane passing through the line $y = x$. Subplot: zoom close to the discontinuity

where

$$U = \begin{pmatrix} \rho \\ \rho v \\ \rho w \\ E \end{pmatrix}, \quad f(U) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ \rho vw \\ v(E+p) \end{pmatrix}, \quad g(U) = \begin{pmatrix} \rho w \\ \rho vw \\ \rho w^2 + p \\ w(E+p) \end{pmatrix}.$$

Here, ρ is the density; v, w are the components of the velocity in the x and y directions; E, the total energy per unit volume; p, the pressure. We consider the equation of state

$$p(\rho, v, w, E) = (\gamma - 1)(E - \frac{\rho}{2}(v^2 + w^2)), \tag{3.3.18}$$

and γ is the ratio of specific heat capacities of the gas taken as 1.4.

We solve numerically (3.3.17) using ACAT2 and ACAT4 for two of the nineteen configurations of the 2-D Riemann problems presented in [78] whose initial conditions are given in Table 3.11. These initial conditions consist of constant states at every quadrant of the spatial domain that are chosen so that the 1D Riemann problems corresponding to two adjacent states consist of only one one-dimensional simple wave: a shock S, a rarefaction wave R, or a slip line i.e. a contact discontinuity with discontinuous tangential velocity J. The sub-indexes $(l, r) \in \{(2,1), (3,2), (3,4), (4,1)\}$ indicate the involved quadrants. For shocks and *rarefaction waves* an over-arrow indicate the direction (backward or forward). And for contact discontinuities a sign $+/-$ is used (instead of the over-arrow), to denote whether it

is a positive or negative slip line.

Test 1	Lax configuration 6			
$p_2 = 1.0$	$\rho_2 = 2.0$	$p_1 = 1.0$	$\rho_1 = 1.0$	
$v_2 = 0.75$	$w_2 = 0.5$	$v_1 = 0.75$	$w_1 = -0.5$	$J^-_{2,1}$
$p_3 = 1.0$	$\rho_3 = 1.0$	$p_4 = 1.0$	$\rho_4 = 3.0$	$J^+_{3,2}$ $J^+_{4,1}$
$v_3 = -0.75$	$w_3 = 0.5$	$v_4 = -0.75$	$w_4 = -0.5$	$J^-_{3,4}$
Test 2	Lax configuration 8			
$p_2 = 1.0$	$\rho_2 = 1.0$	$p_1 = 0.4$	$\rho_1 = 0.5197$	
$v_2 = -0.6259$	$w_2 = 0.1$	$v_1 = 0.1$	$w_1 = 0.1$	$\overleftarrow{R}_{2,1}$
$p_3 = 1.0$	$\rho_3 = 0.8$	$p_4 = 1.0$	$\rho_4 = 1.0$	$J^-_{3,2}$ $\overleftarrow{R}_{4,1}$
$v_3 = 0.1$	$w_3 = 0.1$	$v_4 = 0.1$	$w_4 = -0.6259$	$J^-_{3,4}$

Table 3.11: 2D Euler equations: initial conditions.

These Riemann problems are numerically solved using a (400×400)-point grid and free boundary conditions. The CFL condition used to set the time steps is the following

$$\Delta t = \text{CFL} \min\left(\frac{\Delta x}{s_x^{\max}}, \frac{\Delta y}{s_y^{\max}}\right),$$

where

$$s_x^{\max} = \max_{i,j}\{\left|v_{i,j}^n\right| + c_{i,j}\}, \quad s_y^{\max} = \max_{i,j}\{\left|w_{i,j}^n\right| + c_{i,j}\},$$

with

$$c = \sqrt{\frac{\gamma p}{\rho}}.$$

The CFL parameter is set to 0.4 and time simulation $t = \{0.3, 0.25\}$ respectively.

Figures 3.3.3 and 3.3.5 show the numerical solutions for the density given by ACAT2, ACAT4, WENO3-RK3, and WENO5-RK5. In addition for each case the smoothness indicators ψ_x^1, ψ_y^1, ψ_x^2, ψ_y^2 are shown in figures 3.3.4 and 3.3.6. In all cases, the solutions are stable and similar to those obtained in [74] with a finite volume method. However, the computational cost increases with the order more than for 1D problems.

Tables 3.12-3.13 show the error in L^1-norm corresponding to the solutions provided by ACAT2, ACAT4, WENO3-RK3 and WENO5-RK3 methods. The reference solution is computed using a 1600×1600-point mesh and CFL$= 0.4$.

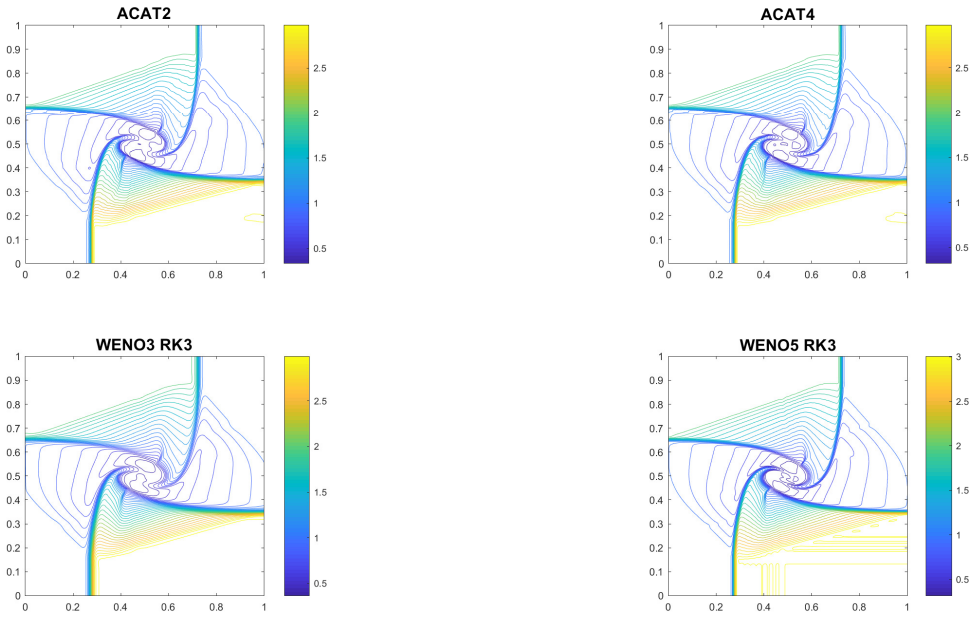

Figure 3.3.3: 2D Euler equations: Lax configuration 6. Contour plots of the density at time $t = 0.3$ obtained with ACAT2 (*left-top*), ACAT4 (*right-top*), WENO3-R3 (*left-bottom*) and WENO5-R3 (*right-bottom*)

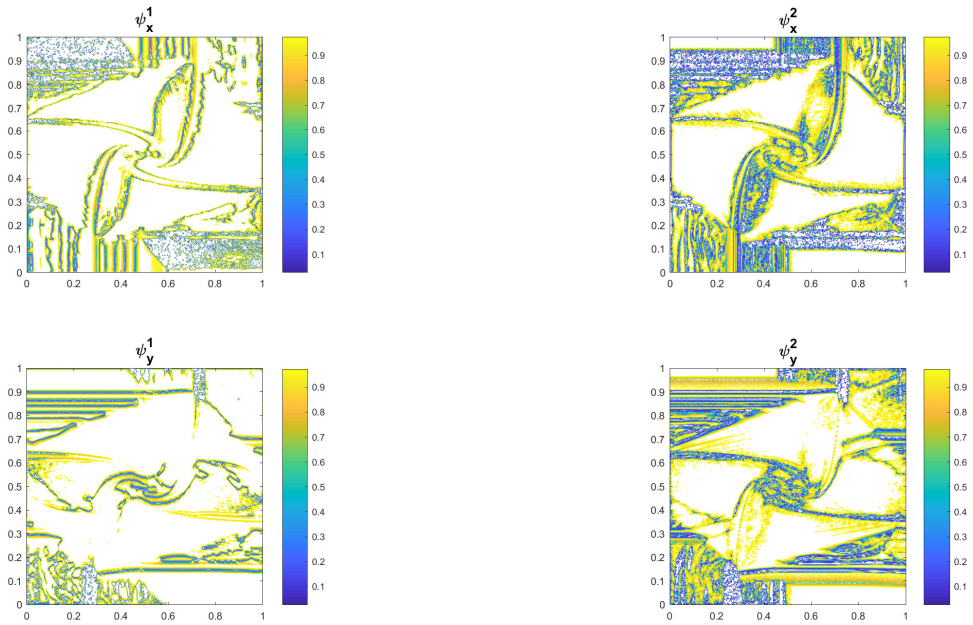

Figure 3.3.4: 2D Euler equations: Lax configuration 6. Contour plots of the smoothness indicators for ACAT2 and ACAT4. ψ_x^1 and ψ_y^1 (*left*). ψ_x^2 and ψ_y^2 (*right*).

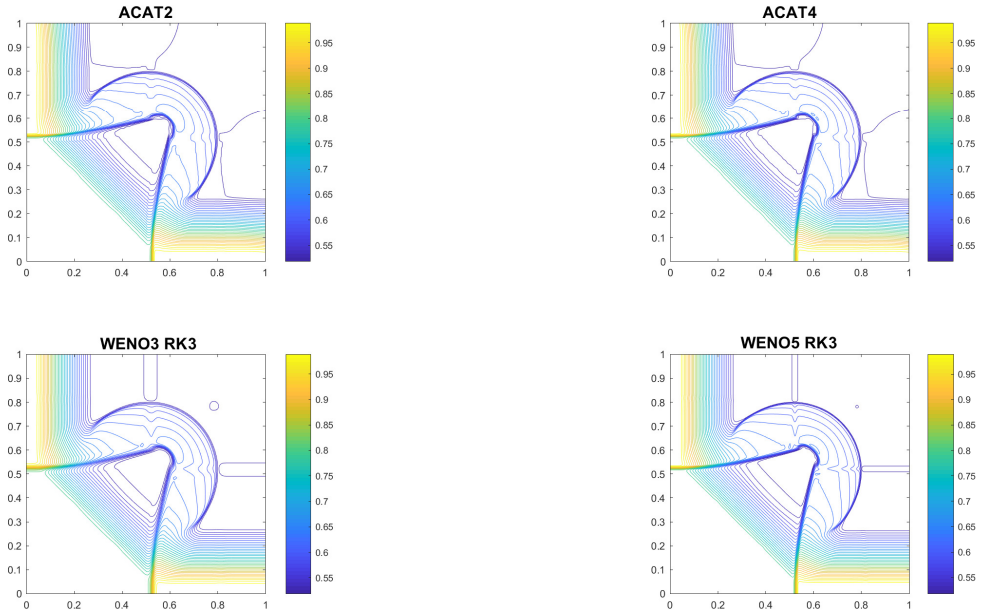

Figure 3.3.5: 2D Euler equations: Lax configuration 8. Contour plots of the density at time $t = 0.25$ obtained with ACAT2 (*left-top*), ACAT4 (*right-top*), WENO3-R3 (*left-bottom*) and WENO5-R3 (*right-bottom*)

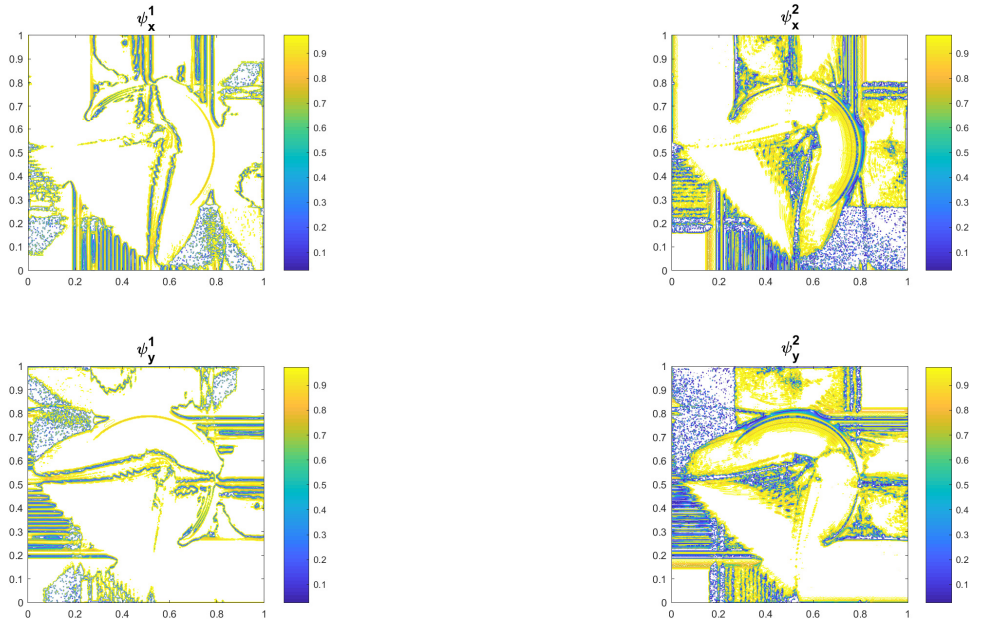

Figure 3.3.6: 2D Euler equations: Lax configuration 8.Contour plots of the smoothness indicators for ACAT2 and ACAT4. ψ_x^1 and ψ_y^1 (*left*). ψ_x^2 and ψ_y^2 (*right*).

	WENO3-RK3				ACAT2			
N	ρ	ρv	ρw	E	ρ	ρv	ρw	E
50	9.8e-2	9.3e-2	6.9e-2	9.1e-2	1.1e-1	1.1e-1	8.0e-2	8.9e-2
100	6.3e-2	6.0e-2	4.3e-2	6.2e-2	7.7e-2	7.7e-2	5.2e-2	6.8e-2
200	3.7e-2	3.2e-2	2.6e-2	3.8e-2	4.8e-2	4.6e-2	3.2e-2	4.0e-2
400	2.2e-2	1.6e-2	1.4e-2	2.1e-2	2.8e-2	2.6e-2	2.0e-2	2.7e-2

Table 3.12: 2D Euler equations: Lax configuration 6. Errors in L^1-norm for ρ, ρv, ρw and E, using CFL= 0.4 and $t = 0.3$.

	WENO5-RK3				ACAT4			
	ρ	ρv	ρw	E	ρ	ρv	ρw	E
50	8.1e-2	8.0e-2	5.9e-2	7.0e-2	7.8e-2	8.3e-2	5.9e-2	7.1e-2
100	5.0e-2	4.9e-2	3.5e-2	5.2e-2	6.0e-2	5.3e-2	5.0e-2	3.3e-2
200	2.8e-2	2.6e-2	2.3e-2	3.2e-2	3.8e-2	2.4e-2	2.6e-2	2.0e-2
400	1.4e-2	1.3e-2	1.8e-2	1.7e-2	2.0e-2	1.3e-2	1.2e-2	1.6e-2

Table 3.13: 2D Euler equations: Lax configuration 6. Errors in L^1-norm for ρ, ρv, ρw and E, using CFL= 0.4 and $t = 0.3$.

Chapter 4

Adaptive Compact Approximate Taylor Method for systems of balance laws and well-balanced property

This chapter is designed to extend the ACAT scheme for hyperbolic nonlinear systems of balance laws

$$U_t + f(U)_x = S(U)H_x, \tag{4.0.1}$$

with initial condition $U(x,0) = U_0(x)$, where $U : \mathbb{R} \times [0, +\infty) \to \mathbb{R}^d$; $f : \mathbb{R}^d \to \mathbb{R}^d$ is the flux function; $S : \mathbb{R}^d \to \mathbb{R}^d$ is the source term; and $H : \mathbb{R} \to \mathbb{R}$ is a known function. PDE systems of this form appear in many fluid models in different contexts: shallow water models, multiphase flow models, gas dynamic, elastic wave equations, etc.

More precisely, we focus on the extension of high-order Lax-Wendroff methods to systems (4.0.1).

Following the strategy in [45] (see also [13, 34]), (4.0.1) is first written in conservative form through the definition of a 'combined flux' formed by the sum of flux function f and the indefinite integral of the source term: more precisely, let us introduce the function $\mathcal{F}$ given by

$$\mathcal{F}(U)(x,t) = f(U(x,t)) - \int_{-\infty}^{x} S(U(\sigma,t))H_x(\sigma)\, d\sigma, \tag{4.0.2}$$

assuming that the integral is finite. Then, the equality

$$\mathcal{F}(U)_x = f(U)_x - S(U)H_x,$$

allows one to write the system of balance laws (4.0.1) in the form

$$U_t + \mathcal{F}(U)_x = 0. \tag{4.0.3}$$

4.1 Compact Approximate Taylor Method for Balance Law

The formal expression of CAT2P for systems of balance law (4.0.3) is given by:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(\mathfrak{F}^P_{i-\frac{1}{2}} - \mathfrak{F}^P_{i+\frac{1}{2}}\Big), \tag{4.1.1}$$

where,

$$\mathfrak{F}^P_{i+\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}\Big(\mathcal{F}^{(k-1)}_{i,*}\Big). \tag{4.1.2}$$

Here, $\mathcal{F}^{(k)}_{i,j}$ are local approximation of $\partial_t^{(k)}\mathcal{F}(U)(x_{i+j}, t_n)$ that are computed by adapting the algorithm described in Section 3.1 to system (4.0.3). Formally, the algorithm is as follows:

- Define

$$\begin{aligned} F^{(0)}_{i,j} &:= f(U^n_{i+j}), \quad j = -P+1, \dots, P; \\ I^{(0)}_{i,j} &:= \int_{-\infty}^{x_{i+j}} S(U(x,t_n))H_x(x)\,dx, \quad j = -P+1, \dots, P. \end{aligned}$$

- For $k = 1, \dots, 2P-1$:

  - Compute for all $j = -P+1, \dots, P$

$$U^{(k)}_{i,j} = -A_P^{1,j}\Big(F^{(k-1)}_{i,*}, \Delta x\Big) + A_P^{1,j}\Big(I^{(k-1)}_{i,*}, \Delta x\Big).$$

  - Define for all $j, r = -P+1, \dots, P$

$$\begin{aligned} I^{n+r}_{i,j} &:= \int_{-\infty}^{x_{i+j}} S(U(x,t_{n+r}))H_x(x)\,dx, \\ F^{k,n+r}_{i,j} &:= f\Big(U^{k,n+r}_{i,j}\Big), \end{aligned}$$

    where $U^{k,n+r}_{i,j}$ is the approximation of $U(x_{i+j}, t_{n+r})$ given by the Taylor expansion

in time:

$$U_{i,j}^{k,n+r} = U_{i+j}^{n} + \sum_{m=1}^{k} \frac{(\Delta t)^m}{m!} U_{i,j}^{(m)}.$$

– Compute for all $j = -P+1, \ldots, P$

$$F_{i,j}^{(k)} = A_P^{k,0}\Big(F_{i,j}^{k,*}, \Delta t\Big), \quad I_{i,j}^{(k)} = A_P^{k,0}\Big(I_{i,j}^{*}, \Delta t\Big).$$

The 'numerical fluxes' are then defined by:

$$\mathfrak{F}_{i+\frac{1}{2}}^{P} = F_{i+\frac{1}{2}}^{P} - I_{i+\frac{1}{2}}^{P} \tag{4.1.3}$$

where

$$F_{i+\frac{1}{2}}^{P} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}(F_{i,*}^{(k-1)}, \Delta x), \tag{4.1.4}$$

$$I_{i+\frac{1}{2}}^{P} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}(I_{i,*}^{(k-1)}, \Delta x). \tag{4.1.5}$$

4.1.1 CAT2P for balance law

The CAT2P algorithm for balance laws, shown previously, is formal; since it requires the computation of integrals that depend on the exact solution in intervals of the form $(-\infty, x_{i+j}]$. In order to be computationally implementable, let us first rewrite it using only integrals in

bounded intervals. To do that, the key point is the following chain of equalities:

$$\begin{aligned}
A_P^{1,j}\Big(I_{i,*}^{(k-1)},\Delta x\Big) &= \frac{1}{\Delta x}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}I_{i,s}^{(k-1)}\\
&= \frac{1}{\Delta x}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}A_P^{k-1,0}\Big(I_{i,s}^{*},\Delta t\Big)\\
&= \frac{1}{\Delta x\Delta t^{k-1}}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}I_{i,s}^{k,n+r}\\
&= \frac{1}{\Delta x\Delta t^{k-1}}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}\int_{-\infty}^{x_{i+s}}S(U(x,t_{n+r}))H_x(x)\,dx\\
&= \frac{1}{\Delta x\Delta t^{k-1}}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}\Big(\int_{-\infty}^{x_{i+s}}S(U(x,t_{n+r}))H_x(x)\,dx\\
&\qquad -\int_{-\infty}^{x_{i-P+1}}S(U(x,t_{n+r}))H_x(x)\,dx\Big)\\
&= \frac{1}{\Delta x\Delta t^{k-1}}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}\int_{x_{i-P+1}}^{x_{i+s}}S(U(x,t_{n+r}))H_x(x)\,dx
\end{aligned}$$

where the equality

$$\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}=0,\quad j=-P+1,\dots,P,$$

has been used. Remember that, an interpolatory formula of numerical differentiation that uses $2P$ points is exact at least for polynomials of degree $2P-1$ and thus it is exact for constant polynomials. Therefore, if the formula is applied to the constant polynomial $p\equiv 1$, we have

$$0=p'(x_{i+j})=A_P^{1,j}(p,\Delta x)=\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}p(x_{i+s})=\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}.$$

By introducing the notation

$$\begin{aligned}
I_{i,j,l}^{m} &:= \int_{x_{i+j}}^{x_{i+l}}S(U(x,t_m))H_x(x)\,dx,\\
I_{i,j,l}^{(k-1)} &:= A_P^{k-1,0}\Big(I_{i,j,l}^{*},\Delta t\Big),
\end{aligned}$$

we obtain

$$\begin{aligned} A_P^{1,j}\Big(I_{i,*}^{(k-1)},\Delta x\Big) &= \frac{1}{\Delta x\Delta t^{k-1}}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}I_{i,-P+1,s}^{n+r} \\ &= \frac{1}{\Delta x}\sum_{r=-P+1}^{P}\gamma_{P,r}^{k-1,0}I_{i,-P+1,r}^{(k-1)} \\ &= A_P^{1,j}\Big(I_{i,-P+1,*}^{(k-1)},\Delta x\Big), \end{aligned}$$

where only integrals in intervals of the form $[x_{i-P+1},x_{i+s}]$ appear for all $s=-P+1,\dots,P$. Observe that $I_{i-P+1,i-P+1}^{(k)}=0$ for all k.

Concerning the expression of the numerical method, observe that:

$$\begin{aligned} I_{i+\frac{1}{2}}^P - I_{i-\frac{1}{2}}^P &= \sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}\left(A_P^{0,1/2}(I_{i,*}^{(k-1)},\Delta x)-A_P^{0,1/2}(I_{i-1,*}^{(k-1)},\Delta x)\right) \\ &= \sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,1/2}\left(I_{i,j}^{(k-1)}-I_{i-1,j}^{(k-1)}\right) \\ &= \sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,1/2}A_P^{k-1,0}\left(I_{i,j}^{*}-I_{i-1,j}^{*},\Delta t)\right). \end{aligned}$$

Since

$$I_{i,j}^{n+r}-I_{i-1,j}^{n+r}=I_{i,j-1,j}^{n+r}=\int_{x_{i+j-1}}^{x_{i+j}}S(U(x,t_{n+r}))H_x(x)\,dx, \tag{4.1.6}$$

if we define

$$\mathcal{I}_{i,j}^{(k-1)}=A_P^{k-1,0}\Big(I_{i,j-1,j}^{*},\Delta t\Big), \tag{4.1.7}$$

we have

$$I_{i+\frac{1}{2}}^P - I_{i-\frac{1}{2}}^P=\sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}A_P^{0,1/2}(\mathcal{I}_{i,*}^{(k-1)},\Delta x),$$

so that (4.1.1) can be written in equivalent form

$$U_i^{n+1}=U_i^n+\frac{\Delta t}{\Delta x}\Big(F_{i-\frac{1}{2}}^P-F_{i+\frac{1}{2}}^P+S_i^P\Big), \tag{4.1.8}$$

where

$$S_i^P=\sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}A_P^{0,1/2}(\mathcal{I}_{i,*}^{(k-1)},\Delta x). \tag{4.1.9}$$

Observe that only integrals (4.1.6) in intervals of length Δx appear in the expression of the

numerical source term.

Quadrature formulas

Furthermore, in order to have an implementable algorithm, all the integrals appearing in source term (4.1.9) might be approximated using quadrature formulas combined with the approximations $U_{i,j}^{k,n+r}$ of the exact solution that are available at every stage. To do this, given i and $j = -P+2, \dots, P$, we consider at $[x_{i+j-1}, x_{i+j}]$ the interpolatory quadrature formula

$$\int_{x_{i+j-1}}^{x_{i+j}} f(x)\,dx \approx \Delta x \sum_{s=-P+1}^{P} a_{P,s}^{i,j} f(x_{i+s})$$

whose nodes are x_{i+s}, $s = -P+1, \dots, P$. This formula will be used to approximate the integrals appearing at the $k-$th stage of the algorithm as follows: given two indices $j_1 < j_2$

$$I_{i,j_1,j_2}^{m} \approx \widetilde{I}_{i,j_1,j_2}^{k,m} := \Delta x \sum_{s=j_1+1}^{j_2} \sum_{l=-P+1}^{P} a_{P,l}^{i,s} S(U_{i,l}^{k,m}) H_x(x_{i+l}).$$

Taking into account these approximations of the integral terms, the algorithm is finally as follows:

- Compute

$$\begin{aligned}
&F_{i,j}^{(0)} = f(U_{i+j}^{n}), \quad j = -P+1, \dots, P;\\
&\widetilde{I}_{i,j-1,j}^{(0)} = \Delta x \sum_{l=-P+1}^{P} a_{P,l}^{i,j} S(U_{i+l}^{n}) H_x(x_{i+l}), \quad j = -P+2, \dots, P;\\
&\widetilde{I}_{i,-P+1,-P+1}^{(0)} = 0;\\
&\widetilde{I}_{i,-P+1,j}^{(0)} = \sum_{s=-P+2}^{j} \widetilde{I}_{i,s-1,s}^{(0)}, \quad j = -P+2, \dots, P.
\end{aligned}$$

- For $k = 1, \dots, 2P-1$:

  - Compute for all $j = -P+1, \dots, P$

$$U_{i,j}^{(k)} = -A_P^{1,j}\Big(F_{i,*}^{(k-1)}, \Delta x\Big) + A_P^{1,j}\Big(I_{i,-P+1,*}^{(k-1)}, \Delta x\Big).$$

– Compute for all $j, r = -P+1, \dots, P$

$$U_{i,j}^{k,n+r} = U_{i+j}^{n} + \sum_{m=1}^{k} \frac{(\Delta t)^m}{m!} U_{i,j}^{(m)}.$$

– Compute for all $j, r = -P+1, \dots, P$

$$F_{i,j}^{k,n+r} = f\Big(U_{i,j}^{k,n+r}\Big),$$

– Compute for all $r = -P+1, \dots, P$, $j = -P+2, \dots, P$

$$\widetilde{I}_{i,j-1,j}^{k,n+r} = \Delta x \sum_{l=-P+1}^{P} a_{P,l}^{i,j} S(U_{i,l}^{k,n+r}) H_x(x_{i+l}).$$

– Compute for all $j = -P+2, \dots, P$

$$\widetilde{I}_{i,j-1,j}^{(k)} = A_P^{k,0}\Big(\widetilde{I}_{i,j-1,j}^{k,*}, \Delta t\Big).$$

– Compute

$$\begin{aligned}
&F_{i,j}^{(k)} = A_P^{k,0}\Big(F_{i,j}^{k,*}, \Delta t\Big), \quad j = -P+1, \dots, P;\\
&\widetilde{I}_{i,-P+1,-P+1}^{(k)} = 0;\\
&\widetilde{I}_{i,-P+1,j}^{(k)} = \sum_{s=-P+2}^{j} \widetilde{I}_{i,s-1,s}^{(k)} \quad j = -P+2, \dots, P.
\end{aligned}$$

Once the algorithm has been executed, the integrals already computed can be used to approximate the source term as follows:

- For $k = 1, \dots, 2P$ define

$$\widetilde{\mathcal{I}}_{i,j}^{(k-1)} = \begin{cases} \widetilde{I}_{i-1,j,j+1}^{(k-1)} & \text{if } j = -P+1, \dots, 0; \\ \\ \widetilde{I}_{i,j-1,j}^{(k-1)} & \text{if } j = 1, \dots, P. \end{cases}$$

- Compute

$$\widetilde{S}_i^P = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}(\widetilde{\mathcal{I}}_{i,*}^{(k-1)}, \Delta x).$$

Observe that the first P integral terms $\mathcal{I}_{i,j}^{(k-1)}$ appearing in the expression of the numerical source term (4.1.9) are approximated with the values $I_{i-1,j,j+1}^{(k-1)}$, used to compute the flux at the intercell $i-\frac{1}{2}$, and the P last ones by $I_{i,j-1,j}^{(k-1)}$, used to compute the flux at the intercell $i+\frac{1}{2}$.

The final expression of the numerical method is then

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F_{i-\frac{1}{2}}^P - F_{i+\frac{1}{2}}^P + \widetilde{S}_i^P\Big), \tag{4.1.10}$$

where $F_{i+\frac{1}{2}}^P$ is given by (4.1.4).

4.1.2 CAT2 for balance law

Let us illustrate the above numerical method in the easiest case $P = 1$. In this case, the quadrature formula used to compute integrals in intervals of length Δx is the trapezoidal rule:

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \frac{\Delta x}{2}\Big(f(x_i) + f(x_{i+1})\Big).$$

The numerical method is then as follows: for every i

- Compute

$$U_{i,j}^{(1)} = -\frac{1}{\Delta x}\left(f(U_{i+1}^n) - f(U_i^n)\right) + \frac{1}{2}\left(S(U_i^n)H_x(x_i) + S(U_{i+1}^n)H_x(x_{i+1})\right), \quad j = 0, 1.$$

- Compute

$$U_{i,j}^{1,n+1} = U_{i+j}^n + \Delta t\, U_{i,j}^{(1)}, \quad j = 0, 1.$$

Then, define

$$F_{i+\frac{1}{2}}^1 := \frac{1}{4}\left(f(U_i^n) + f(U_{i+1}^n) + f(U_{i,0}^{1,n+1}) + f(U_{i,1}^{1,n+1})\right) \tag{4.1.11}$$

$$\begin{aligned}\widetilde{S}_i^1 \;:=\; & \frac{\Delta x}{8}\Big((S(U_{i-1}^n) + S(U_{i-1,0}^{1,n+1}))H_x(x_{i-1}) + (S(U_i^n) + S(U_{i-1,1}^{1,n+1}))H_x(x_i) \\ & +(S(U_i^n) + S(U_{i,0}^{1,n+1}))H_x(x_i) + (S(U_{i+1}^n) + S(U_{i,1}^{1,n+1}))H_x(x_{i+1})\Big).\end{aligned} \tag{4.1.12}$$

The numerical method is then (4.1.10) with $P = 1$.

4.2 Adaptive Compact Approximate Taylor Method for Balance Law

As we have seen, the Compact Approximate Taylor (CAT) schemes (3.1.1) for systems of conservation laws are linearly stable in the L^2-sense under the usual CFL condition. Unfortunately, spurious oscillations may appear close to a discontinuity of the solution, as it happens for the Lax-Wendroff method: see [15]. Then, following the same idea of Chapter 3, we use an extended version of the shock-capturing technique introduced in [14] based on a family of high-order smoothness indicators adapted to systems of balance laws. The idea is as follows: once the approximations at time t^n have been computed, the candidate stencils to compute $F^p_{i+\frac{1}{2}}$ are

$$S^p_{i+\frac{1}{2}} = \{x_{i-p+1}, \ldots, x_{i+p}\}, \quad p = 1, \ldots, P.$$

The selected stencil is the one with maximal length among those in which the solution at time t^n is smooth, according to some smoothness indicators $\psi^p_{i+\frac{1}{2}}$ for $p = 1, \ldots, P$. If a discontinuity is detected in the stencil $S^1_{i+\frac{1}{2}}$ a robust first-order numerical method is used. The ingredients of this strategy are described below: a robust first order method; a family of smoothness indicators and CAT2P scheme for systems of balance laws.

First-order numerical method

As first-order robust scheme to be combined with CAT2P methods for balance laws, we select the Lax-Friedrichs method applied to (4.0.3) what leads to the formal expression:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F^{LF}_{i-\frac{1}{2}} - F^{LF}_{i+\frac{1}{2}} + \widetilde{S}_i^{LF}\Big) \tag{4.2.1}$$

where

$$F^{LF}_{i+\frac{1}{2}} = \frac{1}{2}\left(f(U_i^n) + f(U_{i+1}^n)\right) - \frac{\Delta x}{2\Delta t}\left(U_{i+1}^n - U_i^n\right), \tag{4.2.2}$$

$$\widetilde{S}_i^{LF} = \Delta x S(U_i^n) H_x(x_i), \tag{4.2.3}$$

in which the mid-point rule has been used to approximate the integral corresponding to the source term.

ACAT2 for balance law

The expression of the ACAT2 or Flux Limiter numerical method is based on a flux limiter (see [83, 84, 120]). Its expression is as follows:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F^*_{i-\frac{1}{2}} - F^*_{i+\frac{1}{2}} + \widetilde{S}^*_i\Big) \tag{4.2.4}$$

where

$$F^*_{i+\frac{1}{2}} = \varphi_i^1\, F^1_{i+\frac{1}{2}} + (1-\varphi_i^1)\, F^{LF}_{i+\frac{1}{2}}, \tag{4.2.5}$$

$$\widetilde{S}^*_i = \varphi_i^1 \widetilde{S}^1_i + (1-\varphi_i^1)\widetilde{S}^{LF}_i, \tag{4.2.6}$$

where $F^1_{i+\frac{1}{2}}$ and $\widetilde{S}^1_i$ are given by (4.1.11) and (4.1.12) respectively; $F^{LF}_{i+\frac{1}{2}}$ and $\widetilde{S}^{LF}_i$ are given by (4.2.2) and (4.2.3) respectively;

$$\varphi_i^1 = \min(\varphi^1_{i-\frac{1}{2}}, \varphi^1_{i+\frac{1}{2}}),$$

where $\varphi^1_{i+\frac{1}{2}}$ is a flux limiter, i.e.

$$\varphi^1_{i+\frac{1}{2}} \approx \begin{cases} 1 & \text{if } \{U^n_{i-1}, \ldots, U^n_{i+2}\} \text{ are 'smooth';} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2.7}$$

Remark 4.2.1 *Unfortunately, using a* $3-$*point stencil, is not possible to distinguish between critical point and discontinuity. For this reason, on equation* (4.2.7)*, with smooth we mean smooth without critical point, (see [2, 3, 70, 116]).*

ACAT2P for balance law

The final expression of the adaptive CAT2P method for systems of balance laws is as follows:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F^{\mathcal{A}_i}_{i,i-\frac{1}{2}} - F^{\mathcal{A}_i}_{i,i+\frac{1}{2}} + \widetilde{S}^{\mathcal{A}_i}_i\Big), \tag{4.2.8}$$

where

$$F^{A}_{i,i+\frac{1}{2}} = \begin{cases} F^*_{i,i+\frac{1}{2}} & \text{if } \mathcal{A}_i = \emptyset; \\ F^{p_s}_{i,i+\frac{1}{2}} & \text{where } p_s = \max(\mathcal{A}_i) \text{ otherwise;} \end{cases} \tag{4.2.9}$$

and

$$\widetilde{S}_i^A = \begin{cases} \widetilde{S}_i^* & \text{if } \mathcal{A}_i = \emptyset; \\ \widetilde{S}_i^{p_s} & \text{where } p_s = \max(\mathcal{A}_i) \text{ otherwise.} \end{cases} \tag{4.2.10}$$

Here, $\mathcal{A}_i$ is the set of indices given by

$$\mathcal{A}_i = \{p \in \{2, \ldots, P\} \text{ such that } \psi^{\mathrm{p}}_{\mathrm{i}-\frac{1}{2}} \approx 1 \text{ and } \psi^{\mathrm{p}}_{\mathrm{i}+\frac{1}{2}} \approx 1\}; \tag{4.2.11}$$

$F^*_{i+\frac{1}{2}}$ and $\widetilde{S}_i^*$ are the ACAT2 numerical flux and source terms given by (4.2.5), (4.2.6); $F^{p_s}_{i+\frac{1}{2}}$ and $\widetilde{S}_i^{p_s}$ are the ACAT2p_s numerical fluxes and source terms defined in (4.1.4), (4.1.9).

4.3 Well Balanced Compact Approximate Taylor Method for Balance Law

Systems of balance laws (4.0.1) have non-trivial stationary solutions that satisfy the ODE system

$$f(U)_x = S(U)H_x. \tag{4.3.1}$$

The objective of well balanced schemes is to preserve exactly or with machine precision some of these steady state solutions. For instance, in the context of Shallow water equations Bermúdez and Vázquez-Cendón introduced in [5] the condition called *C-property*: a scheme is said to satisfy this condition if it preserves the water at rest solutions. Since then, many different numerical methods that satisfy this property have been introduced in the literature: see [10, 19, 20, 54, 106, 129] and their references. In the framework of finite difference methods, high-order schemes that satisfy the C-property were introduced in [16] and [130]: while the former was based on the formal writing of the system in conservative form based on the above mentioned technique, the latter relied on the expression of the source term as a function of variables that are constants for the stationary solutions to be preserved: see [131].

In [97] a general technique to derive high-order well-balanced finite-difference methods for systems of balance laws was developed. The idea, inspired on the general technique for finite volume methods discussed in [21], was as follows: let U_i be the numerical approximation of the solution $U(x_i, t)$ at the node x_i at time t and let U_i^* be the stationary solution satisfying

the Cauchy problem:

$$\begin{cases} f(U_i^*)_x = S(U_i^*)H_x, \\ \\ U_i^*(x_i) = U_i. \end{cases} \tag{4.3.2}$$

Then, if U_i^* can be found, one has trivially

$$S(U_i)H_x(x_i) = S(U_i^*(x_i))H_x(x_i) = f(U_i^*(x_i))_x. \tag{4.3.3}$$

Therefore, locally the system of balance laws can be written in conservation form as follows

$$U_t + (f(U) - f(U_i^*(x)))_x = 0.$$

For this reason the goal of this section is to derive a well-balanced version of the CAT2P methods introduced previously following the strategy above explained. The idea is as follows: let us suppose that the initial condition is given by

$$U(x,0) = U^*(x),$$

where U^* is a stationary solution of (4.0.1). Let us introduce then the function $\tilde{\mathcal{F}}$ given by

$$\begin{aligned} \tilde{\mathcal{F}}(U)(x,t) =& \mathcal{F}(U)(x,t)) - \mathcal{F}(U^*)(x) = \\ =& f(U(x,t)) - f(U^*(x)) - \int_{-\infty}^{x} \Big(S(U(\sigma,t)) - S(U^*(\sigma)) \Big) H_\sigma(\sigma) d\sigma. \end{aligned} \tag{4.3.4}$$

Hence, observing that

$$\tilde{\mathcal{F}}(U)_x = f(U)_x - f(U^*)_x - (S(U) - S(U^*))H_x = f(U)_x - S(U)H_x,$$

the system of balance laws (4.0.1) can be formally written in the form

$$U_t + \tilde{\mathcal{F}}(U)_x = 0. \tag{4.3.5}$$

Since, obviously $\tilde{\mathcal{F}}(U^*) = 0$, a numerical method based on the discretization of this conservative form is expected to exactly preserve U^*.

In practice, this strategy is applied as follows: once the approximation U_i^n has been

obtained, we consider the local stationary solution U_i^* that satisfies

$$U_i^*(x_i) = U_i^n,$$

i.e. U_i^* solves the Cauchy problem

$$\begin{cases} f(U)_x = S(U)H_x \\ U(x_i) = U_i^n. \end{cases} \tag{4.3.6}$$

Let us assume for simplicity that this Cauchy problem has a unique solution that is explicitly known. Then, the system of balance laws is locally rewritten in the form

$$U_t + \tilde{\mathcal{F}}_i(U)_x = 0. \tag{4.3.7}$$

where

$$\begin{aligned} \tilde{\mathcal{F}}_i(U)(x,t) =& \mathcal{F}(U)(x,t)) - \mathcal{F}(U_i^*)(x) = \\ =& f(U(x,t)) - f(U_i^*(x)) - \int_{-\infty}^{x} \Big(S(U(\sigma,t)) - S(U_i^*(\sigma)) \Big) H_x(\sigma)\, d\sigma \end{aligned} \tag{4.3.8}$$

and the CAT2P method is then applied:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(\tilde{\mathfrak{F}}^P_{i;i-\frac{1}{2}} - \tilde{\mathfrak{F}}^P_{i;i+\frac{1}{2}} \Big), \tag{4.3.9}$$

where,

$$\tilde{\mathfrak{F}}^P_{i;i+\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}\Big(\tilde{\mathcal{F}}^{(k-1)}_{i;i,*} \Big), \tag{4.3.10}$$

$$\tilde{\mathfrak{F}}^P_{i;i-\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,1/2}\Big(\tilde{\mathcal{F}}^{(k-1)}_{i;i-1,*} \Big). \tag{4.3.11}$$

Here $\tilde{\mathcal{F}}^{(k)}_{i;l,j}$ is an approximation of $\partial_t^{(k)} \tilde{\mathcal{F}}_i(U)(x_{l+j}, t_n)$.

Remark 4.3.1 *Observe that, in this case, two numerical fluxes have to be computed at every inter-cell $x_{i+1/2}$, $\tilde{\mathfrak{F}}^P_{i;i+\frac{1}{2}}$ and $\tilde{\mathfrak{F}}^P_{i+1;i+\frac{1}{2}}$, whose computation are based respectively on the stationary solutions U_i^* (that satisfies $U_i^*(x_i) = U_i^n$) and U_{i+1}^* (that satisfies $U_{i+1}^*(x_{i+1}) = U_{i+1}^n$).*

Remark 4.3.2 *Observe that, if the initial condition U_0 is a stationary solution, then at time $t = 0$, $U_i^* = U_0$ for all i, so that*

$$\tilde{\mathcal{F}}_i(U_0) = \mathcal{F}(U_0) - \mathcal{F}(U_i^*) = \mathcal{F}(U_0) - \mathcal{F}(U_0) = 0, \quad \forall i,$$

and the numerical method is expected to preserve exactly the initial condition.

4.3.1 WBCAT2 for balance law

In order to simplify the derivation of the well-balanced Compact Approximate Taylor scheme, let us illustrate the second order numerical method, i.e. $P = 1$. For every i :

- Compute the solution U_i^* of the Cauchy problem (4.3.6).

- Compute

$$\begin{aligned}
U_{i;i,j}^{(1)} &= -\frac{1}{\Delta x}\Big(f(U_{i+1}^n) - f(U_i^n) - f(U_i^*(x_{i+1})) + f(U_i^*(x_i))\Big) \\
&\quad +\frac{1}{2}\Big((S(U_i^n) - S(U_i^*(x_i)))H_x(x_i) + (S(U_{i+1}^n) - S(U_i^*(x_{i+1})))H_x(x_{i+1})\Big), \quad j = 0, 1; \\
U_{i;i-1,j}^{(1)} &= -\frac{1}{\Delta x}\Big(f(U_i^n) - f(U_{i-1}^n) - f(U_i^*(x_i)) + f(U_i^*(x_{i-1}))\Big) \\
&\quad +\frac{1}{2}\Big((S(U_{i-1}^n) - S(U_i^*(x_{i-1})))H_x(x_{i-1}) + (S(U_i^n) - S(U_i^*(x_i)))H_x(x_i)\Big), \quad j = 0, 1.
\end{aligned}$$

- Compute

$$\begin{aligned}
U_{i;i,j}^{1,n+1} &= U_{i+j}^n + \Delta t\, U_{i;i,j}^{(1)}, \quad j = 0, 1, \\
U_{i;i-1,j}^{1,n+1} &= U_{i+j}^n + \Delta t\, U_{i-1;i,j}^{(1)}, \quad j = 0, 1.
\end{aligned}$$

- Define

$$
\begin{aligned}
F^1_{i;i+\frac{1}{2}} &:= \frac{1}{4}\left(f(U^n_i)+f(U^n_{i+1})+f(U^{1,n+1}_{i;i,0})+f(U^{1,n+1}_{i;i,1})-2f(U^*_i(x_i))-2f(U^*_i(x_{i+1}))\right), && (4.3.12)\\
F^1_{i;i-\frac{1}{2}} &:= \frac{1}{4}\left(f(U^n_{i-1})+f(U^n_i)+f(U^{1,n+1}_{i;i-1,0})+f(U^{1,n+1}_{i;i-1,1})-2f(U^*_i(x_{i-1}))-2f(U^*_i(x_i))\right), && (4.3.13)\\
\widetilde{S}^1_i &:= \frac{\Delta x}{8}\Big((S(U^n_{i-1})+S(U^{1,n+1}_{i-1,0})-2S(U^*_i(x_{i-1})))H_x(x_{i-1}) && (4.3.14)\\
&\quad +(S(U^n_i)+S(U^{1,n+1}_{i;i-1,1})-2S(U^*_i(x_i)))H_x(x_i)\\
&\quad +(S(U^n_i)+S(U^{1,n+1}_{i;i,0})-2S(U^*_i(x_i)))H_x(x_i)\\
&\quad +(S(U^n_{i+1})+S(U^{1,n+1}_{i;i,1})-2S(U^*_i(x_{i+1})))H_x(x_{i+1})\Big).
\end{aligned}
$$

Hence, the second order numerical method WBCAT2 is so defined:

$$
U^{n+1}_i = U^n_i + \frac{\Delta t}{\Delta x}\Big(F^1_{i;i-\frac{1}{2}} - F^1_{i;i+\frac{1}{2}} + \widetilde{S}^1_i\Big). \tag{4.3.15}
$$

4.3.2 WBCAT2P for balance law

The algorithm for the high order case is as follows. For every i :

- Compute the solution U^*_i of the Cauchy problem (4.3.6).

- Compute

$$
\begin{aligned}
&F^{(0)}_{i;i,j} = f(U^n_{i+j}) - f(U^*_i(x_{i+j})), \quad j=-P+1,\dots,P;\\
&F^{(0)}_{i;i-1,j} = f(U^n_{i-1+j}) - f(U^*_i(x_{i-1+j})), \quad j=-P+1,\dots,P;\\
&\widetilde{I}^{(0)}_{i;i,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i,j}_{P,l}\left(S(U^n_{i+l}) - S(U^*_i(x_{i+l}))\right)H_x(x_{i+l}), \quad j=-P+2,\dots,P;\\
&\widetilde{I}^{(0)}_{i;i,-P+1,-P+1} = 0;\\
&\widetilde{I}^{(0)}_{i;i,-P+1,j} = \sum_{s=-P+2}^{j} \widetilde{I}^{(0)}_{i;i,s-1,s}, \quad j=-P+2,\dots,P;\\
&\widetilde{I}^{(0)}_{i;i-1,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i-1,j}_{P,l}\left(S(U^n_{i-1+l}) - S(U^*_i(x_{i-1+l}))\right)H_x(x_{i-1+l}), \quad j=-P+2,\dots,P;\\
&\widetilde{I}^{(0)}_{i;i-1,-P+1,-P+1} = 0;\\
&\widetilde{I}^{(0)}_{i;i-1,-P+1,j} = \sum_{s=-P+2}^{j} \widetilde{I}^{(0)}_{i;i-1,s-1,s}, \quad j=-P+2,\dots,P.
\end{aligned}
$$

- For $k=1,\dots,2P-1$:

– Compute for all $j = -P+1, \ldots, P$

$$U^{(k)}_{i;i,j} = -A^{1,j}_P\Big(F^{(k-1)}_{i;i,*}, \Delta x\Big) + A^{1,j}_P\Big(I^{(k-1)}_{i;i,-P+1,*}, \Delta x\Big);$$
$$U^{(k)}_{i;i-1,j} = -A^{1,j}_P\Big(F^{(k-1)}_{i;i-1,*}, \Delta x\Big) + A^{1,j}_P\Big(I^{(k-1)}_{i;i-1,-P+1,*}, \Delta x\Big).$$

– Compute for all $j, r = -P+1, \ldots, P$

$$U^{k,n+r}_{i;i,j} = U^n_{i+j} + \sum_{m=1}^{k} \frac{(\Delta t)^m}{m!} U^{(m)}_{i;i,j},$$
$$U^{k,n+r}_{i;i-1,j} = U^n_{i+j-1} + \sum_{m=1}^{k} \frac{(\Delta t)^m}{m!} U^{(m)}_{i;i-1,j}.$$

– Compute for all $j, r = -P+1, \ldots, P$

$$F^{k,n+r}_{i;i,j} = f\Big(U^{k,n+r}_{i;i,j}\Big) \quad \text{and} \quad F^{k,n+r}_{i;i-1,j} = f\Big(U^{k,n+r}_{i;i-1,j}\Big).$$

– Compute for all $r = -P+1, \ldots, P$, $j = -P+2, \ldots, P$

$$\widetilde{I}^{k,n+r}_{i;i,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i,j}_{P,l} S(U^{k,n+r}_{i;i,l}) H_x(x_{i+l}),$$
$$\widetilde{I}^{k,n+r}_{i;i-1,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i-1,j}_{P,l} S(U^{k,n+r}_{i;i-1,l}) H_x(x_{i-1+l}).$$

– Compute for all $j = -P+2, \ldots, P$

$$\widetilde{I}^{(k)}_{i;i,j-1,j} = A^{k,0}_P\Big(\widetilde{I}^{k,*}_{i;i,j-1,j}, \Delta t\Big), \quad \widetilde{I}^{(k)}_{i;i-1,j-1,j} = A^{k,0}_P\Big(\widetilde{I}^{k,*}_{i;i-1,j-1,j}, \Delta t\Big).$$

– Compute

$$
\begin{aligned}
&F^{(k)}_{i;i,j} = A^{k,0}_P\Big(F^{k,*}_{i;i,j}, \Delta t\Big), \quad j = -P+1, \dots, P;\\
&\widetilde{I}^{(k)}_{i;i,-P+1,-P+1} = 0;\\
&\widetilde{I}^{(k)}_{i;i,-P+1,j} = \sum_{s=-P+2}^{j} \widetilde{I}^{(k)}_{i;i,s-1,s} \quad j = -P+2, \dots, P;\\
&F^{(k)}_{i;i-1,j} = A^{k,0}_P\Big(F^{k,*}_{i;i-1,j}, \Delta t\Big), \quad j = -P+1, \dots, P;\\
&\widetilde{I}^{(k)}_{i;i-1,-P+1,-P+1} = 0;\\
&\widetilde{I}^{(k)}_{i;i-1,-P+1,j} = \sum_{s=-P+2}^{j} \widetilde{I}^{(k)}_{i;i-1,s-1,s} \quad j = -P+2, \dots, P.
\end{aligned}
$$

Once the algorithm has been executed, the integrals already computed can be used to approximate the source term as follows:

- For $k = 1, \dots, 2P$ define

$$
\widetilde{\mathcal{I}}^{(k-1)}_{i,j} = \begin{cases} \widetilde{I}^{(k-1)}_{i;i-1,j,j+1} & \text{if } j = -P+1, \dots, 0; \\ \\ \widetilde{I}^{(k-1)}_{i;i,j-1,j} & \text{if } j = 1, \dots, P. \end{cases}
$$

- Compute

$$
\widetilde{S}^P_i = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A^{0,1/2}_P(\widetilde{\mathcal{I}}^{(k-1)}_{i,*}, \Delta x). \tag{4.3.16}
$$

The final expression of the high order numerical method is then

$$
U^{n+1}_i = U^n_i + \frac{\Delta t}{\Delta x}\Big(F^P_{i;i-\frac{1}{2}} - F^P_{i;i+\frac{1}{2}} + \widetilde{S}^P_i\Big), \tag{4.3.17}
$$

where $F^P_{i;i\pm\frac{1}{2}}$ are given by

$$
F^P_{i;i+\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A^{0,1/2}_P(F^{(k-1)}_{i;i,*}, \Delta x), \tag{4.3.18}
$$

$$
F^P_{i;i-\frac{1}{2}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A^{0,1/2}_P(F^{(k-1)}_{i;i-1,*}, \Delta x). \tag{4.3.19}
$$

Remark 4.3.3 *Observe that this algorithm can be used to update the solution at the point x_i at time t_n only if the Cauchy problem* (4.3.6) *has a solution that is defined in the cells of the stencils $S^P_{i\pm\frac{1}{2}}$ whose analytic expression is known. Therefore:*

- *If* (4.3.6) *has no solution, the CAT2P method will be used instead. Please note that this choice does not spoil the well-balanced character of the numerical method: in this case, the cell values in the stencil cannot be the point values of a stationary solution (otherwise there would be at least one solution of* (4.3.6)*) and thus there is no local equilibrium to preserve.*

- *If* (4.3.6) *has more than one solution, a criterion is needed to select one of them: this is the case for the shallow water and Euler with gravity system that will be discussed in Section 4.5.*

- *If* (4.3.6) *has a solution defined in the stencils but it is not possible to find its expression by analytic procedures, it is possible to apply an ODE solver to approximate it, like it has been done in [19] for finite-volume methods. In all the problems considered in Section 4.5 the analytic expression of the stationary solutions is available either in explicit or implicit form.*

Well-balanced property

Numerical method (4.3.17) is fully well-balanced in the following sense:

Theorem 4.3.1 *Let U^* be a continuous stationary solution of* (4.0.1)*. Thus, if the numerical method* (4.3.17) *is applied to the initial condition*

$$U_i^0 = U^*(x_i), \quad \forall i,$$

we obtain

$$U_i^n = U_i^0, \quad \forall i, n.$$

Proof. Observe first that U^* solves any Cauchy problem (4.3.6) for $n = 0$. Therefore, at the first step, the solution of (4.3.6) is given by

$$U_i^* = U^*, \quad \forall i.$$

Therefore, for every i:

$$
\begin{aligned}
&F^{(0)}_{i;i,j} = F^{(0)}_{i;i-1,j} = 0, \quad j = -P+1, \dots, P;\\
&\widetilde{I}^{(0)}_{i;i,j-1,j} = \widetilde{I}^{(0)}_{i;i-1,j-1,j} = 0, \quad j = -P+2, \dots, P;\\
&\widetilde{I}^{(0)}_{i;i,-P+1,j} = \widetilde{I}^{(0)}_{i;i-1,-P+1,j} = 0, \quad j = -P+1, \dots, P;
\end{aligned}
$$

and thus

$$
U^{(1)}_{i;i,j} = U^{(1)}_{i;i-1,j} = 0, \quad j = -P+1, \dots, P.
$$

As a consequence:

$$
\begin{aligned}
&U^{1,r}_{i;i,j} = U^{0}_{i+j}, \quad F^{1,r}_{i;i,j} = f\Big(U^{0}_{i+j}\Big), \quad j, r = -P+1, \dots, P;\\
&U^{1,r}_{i;i-1,j} = U^{0}_{i-1+j}, \quad F^{1,r}_{i;i-1,j} = f\Big(U^{0}_{i-1+j}\Big), \quad j, r = -P+1, \dots, P\\
&\widetilde{I}^{1,r}_{i;i,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i,j}_{P,l} S(U^{0}_{i+l}) H_x(x_{i+l}),\\
&\widetilde{I}^{1,r}_{i;i-1,j-1,j} = \Delta x \sum_{l=-P+1}^{P} a^{i-1,j}_{P,l} S(U^{0}_{i-1+l}) H_x(x_{i-1+l}).
\end{aligned}
$$

Notice that the values of all these quantities do not depend on r. Therefore, when numerical differentiation in time is applied we obtain:

$$
\begin{aligned}
&F^{(1)}_{i;i,j} = F^{(1)}_{i;i-1,j} = 0, \quad j = -P+1, \dots, P;\\
&\widetilde{I}^{(1)}_{i;i,j-1,j} = \widetilde{I}^{(1)}_{i;i-1,j-1,j} = 0, \quad j = -P+2, \dots, P;\\
&\widetilde{I}^{(1)}_{i;i,-P+1,j} = \widetilde{I}^{(1)}_{i;i-1,-P+1,j} = 0, \quad j = -P+1, \dots, P.
\end{aligned}
$$

Therefore

$$
U^{(2)}_{i;i,j} = U^{(2)}_{i;i-1,j} = 0, \quad j = -P+1, \dots, P.
$$

Repeating the reasoning we obtain

$$
\begin{aligned}
&F^{(k)}_{i;i,j} = F^{(k)}_{i;i-1,j} = 0, \quad j = -P+1, \dots, P, \quad k = 0, \dots, 2P;\\
&\widetilde{I}^{(k)}_{i;i,j-1,j} = \widetilde{I}^{(k)}_{i;i-1,j-1,j} = 0, \quad j = -P+2, \dots, P, \quad k = 0, \dots, 2P;\\
&\widetilde{I}^{(1)}_{i;i,-P+1,j} = \widetilde{I}^{(1)}_{i;i-1,-P+1,j} = 0, \quad j = -P+1, \dots, P, \quad k = 0, \dots, 2P.
\end{aligned}
$$

Hence,

$$F^P_{i;i+\frac{1}{2}} = F^P_{i,i-\frac{1}{2}} = S^P_{i+\frac{1}{2}} = 0,$$

with consequence that

$$U^1_i = U^0_i, \quad \forall i,$$

as we wanted to prove. ■

Remark 4.3.4 *The strategy described in Subsection 4.3 should be adapted to obtain schemes that only preserve a specified set of stationary solutions: for instance, this would be the case if the set that must be preserved is a k-parameter family of stationary solutions,*

$$U^*(x; C_1, \ldots, C_k),$$

with $k < d$ and d is the number of variables. If it is the case, instead of looking at a solution of (4.3.6)*, one looks to a solution of the following nonlinear system:*

Find $C^i_1, \ldots, C^i_k$ such that:

$$u^*_{j_l}(x_i; C^i_1, \ldots, C^i_k) = u_{i,j_l}, \quad l = 1, \ldots, k, \tag{4.3.20}$$

where u^*_j, $u_{i,j}$ denote respectively the jth component of U^* and U_i and $\{j_1, \ldots, j_k\}$ is a set of k indices that is predetermined to have the same number of unknowns and equations in (4.3.20)*. These indices $j_1, \ldots, j_k$ are chosen so that systems of equations* (4.3.20) *have a unique solution, when it is possible. Once the problem has been solved, the numerical fluxes and source terms must be computed as in Section 4.3 with the choice*

$$U^*_i(x) = U^*(x, C^i_1, \ldots, C^i_k).$$

4.4 Adaptive Well Balanced Compact Approximate Taylor Method for Balance Law

Unfortunately, the oscillatory behaviour of CAT procedure is maintained also for the well-balanced extension. For this reason the shock-capturing technique introduce above will be employed even in this case.

First-order well-balanced numerical method for balance law

As we have seen, the flux limiter method need a first order numerical scheme to be combined with the high order well-balanced CAT2P methods then a well-balanced version of the Lax-Friedrichs methods is proposed. The formal expression of the well-balanced Lax-Friedrichs method applied to systems (4.3.7) is:

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F_{i-\frac{1}{2}}^{LF} - F_{i+\frac{1}{2}}^{LF} + \widetilde{S}_i^{LF}\Big) \tag{4.4.1}$$

where

$$F_{i+\frac{1}{2}}^{LF} = \frac{1}{2}\left(F(U_i^n) + F(U_{i+1}^n)\right) - \frac{\Delta x}{2\Delta t}\left(U_{i+1}^n - U_i^n\right), \tag{4.4.2}$$

and

$$\widetilde{S}_i^{LF} = F_{i;i+\frac{1}{2}}^{LF,*} - F_{i;i-\frac{1}{2}}^{LF,*}, \tag{4.4.3}$$

in which,

$$F_{i;i+\frac{1}{2}}^{LF,*} = \frac{1}{2}\left(F(U_i^*(x_i)) + F(U_i^*(x_{i+1}))\right) - \frac{\Delta x}{2\Delta t}\left(U_i^*(x_{i+1}) - U_i^*(x_i)\right) \tag{4.4.4}$$

and U_i^* is the stationary solution that satisfies

$$U_i^*(x_i) = U_i^n.$$

(4.4.3) is a consistent approximation of the integral of the source term, since

$$F_{i;i+\frac{1}{2}}^{LF,*} - F_{i;i-\frac{1}{2}}^{LF,*} \approx \Delta x F(U^*)_x(x_i) = \Delta x S(U_i^*(x_i))H_x(x_i) = \Delta x S(U_i^n)H_x(x_i).$$

WBACAT2 for balance law

The well-balanced version of the ACAT2 method has the form

$$U_i^{n+1} = U_i^n + \frac{\Delta t}{\Delta x}\Big(F_{i;i-\frac{1}{2}}^{*} - F_{i;i+\frac{1}{2}}^{*} + \widetilde{S}_i^{*}\Big) \tag{4.4.5}$$

where

$$F^*_{i;i\pm\frac{1}{2}} = \varphi^1_i\, F^1_{i;i\pm\frac{1}{2}} + (1-\varphi^1_i)\, F^{LF}_{i+\frac{1}{2}}, \tag{4.4.6}$$

$$\widetilde{S}^*_i = \varphi^1_i \widetilde{S}^1_i + (1-\varphi^1_i)\widetilde{S}^{LF,*}_i = \varphi^1_i \widetilde{S}^1_i - (1-\varphi^1_i)\left(F^{LF,*}_{i;i+\frac{1}{2}} - F^{LF,*}_{i;i-\frac{1}{2}}\right), \tag{4.4.7}$$

where $F^1_{i;i\pm\frac{1}{2}}$ are given by (4.3.12)-(4.3.13); $F^{LF,*}_{i;i\pm\frac{1}{2}}$ is given by (4.4.4); and $\widetilde{S}^1_i$ is given by (4.3.14).

WBACAT2P for balance law

Analogously to the adaptive strategy to conservation law, let be

$$\mathcal{A}_i = \{p \in \{2,\dots,P\} \; s.t. \; \psi^p_{i-\frac{1}{2}} \approx 1 \text{ and } \psi^p_{i+\frac{1}{2}} \approx 1\}; \tag{4.4.8}$$

then, the expression of the well-balanced adaptive CAT2P method is as follows:

$$U^{n+1}_i = U^n_i + \frac{\Delta t}{\Delta x}\Big(F^{\mathcal{A}_i}_{i;i-\frac{1}{2}} - F^{\mathcal{A}_i}_{i;i+\frac{1}{2}} + \widetilde{S}^{\mathcal{A}_i}_i\Big), \tag{4.4.9}$$

where

$$F^{\mathcal{A}_i}_{i;i+\frac{1}{2}} = \begin{cases} F^*_{i;i+\frac{1}{2}} & \text{if } \mathcal{A}_i = \emptyset; \\ F^{p_s}_{i;i+\frac{1}{2}} & \text{where } p_s = \max(\mathcal{A}_i) \text{ otherwise;} \end{cases} \tag{4.4.10}$$

and

$$\widetilde{S}^{\mathcal{A}_i}_i = \begin{cases} \widetilde{S}^*_i & \text{if } \mathcal{A}_i = \emptyset; \\ \widetilde{S}^{p_s}_i & \text{where } p_s = \max(\mathcal{A}_i) \text{ otherwise;} \end{cases} \tag{4.4.11}$$

where $F^*_{i;i\pm\frac{1}{2}}$ and $\widetilde{S}^*_i$ are the WBACAT2 numerical fluxes and source terms given by (4.4.6), (4.4.7); $F^{p_s}_{i;i+\frac{1}{2}}$ and $\widetilde{S}^{p_s}_i$ are WBCAT2p_s numerical fluxes and source term given by (4.3.19), (4.3.18), (4.3.16) with $P = p_s$.

4.5 Numerical experiments

In this section we apply ACAT2P and WBACAT2P, $P = 1, 2$ methods to several problems: the linear transport equation and Burgers equation with source term, the 1D shallow water system and the 1D Euler equation with gravity. The Minmod flux limiter [105] is used in ACAT2 and the smoothness indicators (3.2.9) are used for ACAT4: no loss of precision for

first order critical points has been observed in any of the test problems considered here due to the use of $\psi^2_{i+\frac{1}{2}}$. Fornberg's algorithm [41, 42] is used to compute the coefficients of the numerical differentiation formulas.

4.5.1 Linear Equation

We consider the linear scalar balance law

$$u_t + u_x = u, \tag{4.5.1}$$

that has the form (4.0.1) with $H(x) = x$. The analytic solution of the initial value problem with condition

$$u(x, 0) = u_0(x)$$

is given by:

$$u(x, t) = u_0(x - t)e^t. \tag{4.5.2}$$

The stationary solutions solve the ODE

$$u_x = u.$$

Hence, the set of stationary solutions is

$$u^*(x) = Ce^x, \quad C \in \mathbb{R}.$$

Order test

Following [97], we consider (4.5.1) with initial condition

$$u_0(x) = \begin{cases} 0 & \text{if} \quad x < 0; \\ p(x) & \text{if} \quad 0 \leq x \leq 1; \\ 1 & \text{if} \quad x > 1; \end{cases} \tag{4.5.3}$$

where $p(x)$ is the polynomial that satisfies $p(0) = 0$, $p(1) = 1$, $p^k(0) = p^k(1) = 0$,

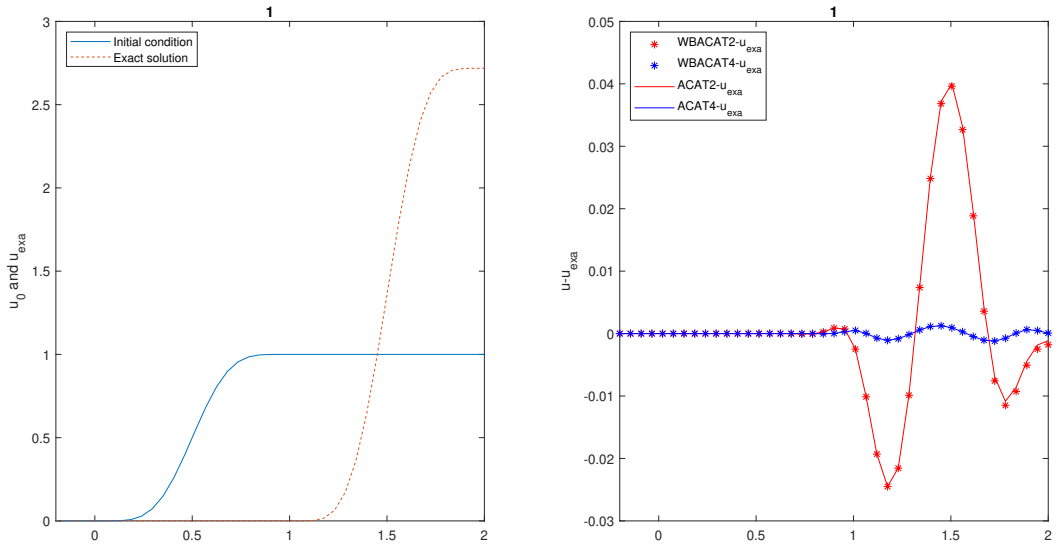

Figure 4.5.1: Test 4.5.1. (Order test). Initial condition and exact solution (left); differences between the numerical solutions and the exact one computed with CAT2, CAT4, WBCAT2 and WBCAT4 at $t = 1$ using a mesh of 41 points (right) on the interval $[-0.2, 2]$ and CFL$= 0.9$.

$k = 1, \ldots, 5$:

$$p(x) = x^6\Big(\sum_{k=0}^{5}(-1)^k\binom{5+k}{k}(x-1)^k\Big)$$

(see Figure 6.4.1). The methods ACAT2, ACAT4, WBACAT2, WBACAT4 have been applied to (4.5.1) with initial condition (4.5.3) in the spatial interval $[-0.2, 2]$, with CFL$= 0.9$. Dirichlet boundary conditions are considered to the left and free boundary conditions to the right based on the use of ghost cells. Figure 4.5.1 shows the numerical solutions obtained at time $t = 1$ on the interval $[-0.2, 2]$ using 40 mesh points.

Tables 4.1-4.2 show the L^1-errors and the empirical order of convergence corresponding to the standard and Adaptive CAT2P and WBCAT2P with $P = 1, 2$. As it can be seen, all the schemes keep the expected order and the errors corresponding to methods of the same order are almost identical. In the first case, the smoothness indicators of ACAT2P and WBACAT2P have been fixed to 1, hence, the Adaptive CAT2P coincides exactly with the standard CAT2P scheme; regarding the second case, no restrictions are imposed on the smoothness indicators, requiring an increase in the number of points to capture the theoretical order. No further restrictions are required for the time step.

	CAT2		WBCAT2		CAT4		WBCAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
6	-	1.79E-1	-	1.80E-1	-	6.03E-2	-	6.03E-2
11	1.57	6.00E-2	1.56	6.05E-2	2.86	8.31E-3	2.86	8.31E-3
21	1.91	1.60E-2	1.92	1.60E-2	3.58	6.94E-4	3.58	6.94E-4
41	2.02	3.93E-3	2.02	3.94E-3	3.88	4.69E-5	3.89	4.69E-5
81	2.03	9.63E-4	2.02	9.66E-4	4.01	2.90E-6	4.00	2.90E-7

Table 4.1: Test 4.5.1: (Order test). Errors in L^1-norm and convergence rates for CAT2, CAT4, WBCAT2 and WBCAT4 at time $t = 1$.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
16	-	2.00E-1	-	1.99E-1	-	1.57E-2	-	1.61E-2
31	1.51	7.01E-2	1.50	7.04E-2	3.01	1.95E-3	2.99	2.02E-3
61	1.85	1,94E-2	1.86	1.94E-2	3.62	1.58E-4	3.60	1.67E-4
121	2.19	4.25E-3	2.19	4.25E-3	8.10	5.71E-7	8.20	5.71E-7
241	2.00	1.05E-3	2.00	1.06E-3	3.99	3.57E-8	4.00	3.56E-8
481	2.00	2.64E-4	2.00	2.64E-4	4.00	2.22E-9	4.00	2.22E-9

Table 4.2: Test 4.5.1: (Order test). Errors in L^1-norm and convergence rates for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 1$.

A moving discontinuity linking two stationary solutions

As a second experiment we consider (4.5.1) with initial condition

$$u_0(x) = \begin{cases} 4e^x & \text{if x} < 0; \\ e^x & \text{otherwise.} \end{cases} \tag{4.5.4}$$

The exact solution consists of a discontinuity linking two stationary solutions that moves with speed 1 in the following way:

$$u(x,t) = \begin{cases} 4e^x & \text{if x} < \text{t}; \\ e^x & \text{otherwise.} \end{cases} \tag{4.5.5}$$

We solve numerically (4.5.1) with initial condition (4.5.4) with ACAT2, ACAT4, WBACAT2, WBACAT4 in the spatial interval $[-\frac{1}{2}, 2]$, using a 100 mesh points, and CFL= 0.9. Dirichlet boundary conditions to the left and free boundary conditions to the right are con-

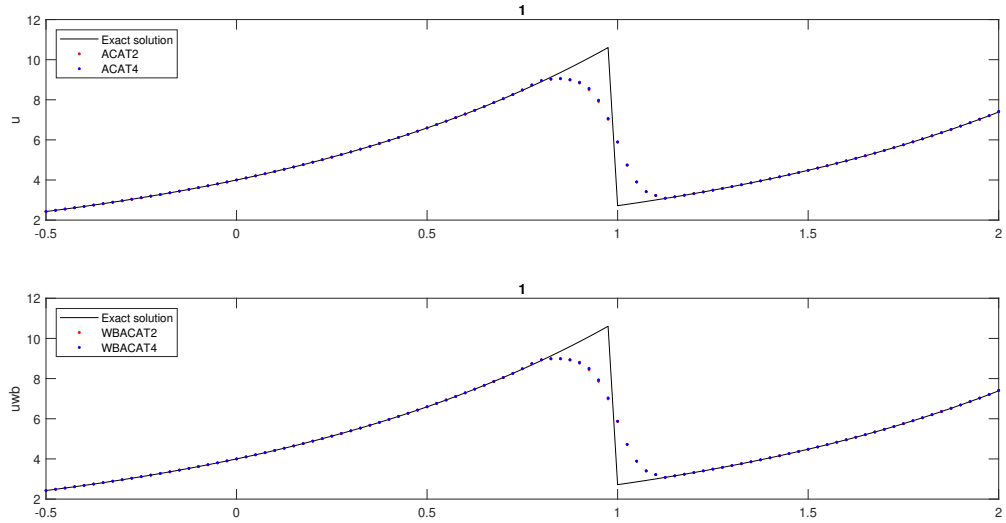

Figure 4.5.2: Test 4.5.1 (A moving discontinuity linking two stationary solutions). Exact and numerical solutions computed with ACAT2-4 (top) and WBACAT2-4 (bottom) at $t = 1$ using a mesh of 100 points.

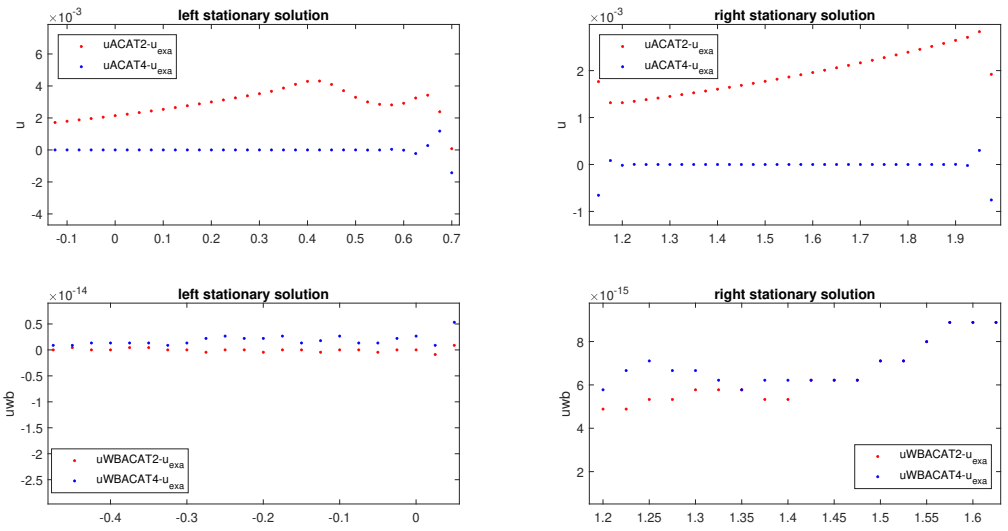

Figure 4.5.3: Test 4.5.1 (A moving discontinuity linking two stationary solutions). Zoom of the left and right differences between the numerical solutions and the exact solution for the no well-balanced (top) and well-balanced (bottom) methods.

sidered.

What we can observe in Figure 4.5.2 is that all the method are able to evolve the discontinuity but, due to the order adaptive nature, they tent to coincide close to the discontinuity and it seems that all the methods are the same. Figure 4.5.3, that shows the differences between the numerical solutions and the exact one, exhibits that, even if they are very similar close to the discontinuity, the error in the smooth region, i.e. the region in which the order is not reduced by the smoothness indicators, is in accordance with the theoretical order. As

expected, far from the discontinuity the well-balanced methods are able to preserve the left and right stationary solution with machine precision.

4.5.2 Burgers Equation

In this section we consider the scalar Burgers equation with source term:

$$u_t + \left(\frac{u^2}{2}\right)_x = u^2 H_x(x). \tag{4.5.6}$$

The stationary solutions solve now the ODE

$$u_x = u H_x(x)$$

whose general solution is

$$u^*(x) = Ce^{H(x)}, \quad C \in \mathbb{R}.$$

Preservation of a stationary solution with linear H

Let us consider $H(x) = x$, so that the stationary solutions become $u^*(x) = Ce^x$. We solve (4.5.6) with initial condition

$$u_0(x) = e^x$$

in the interval $[-1, 1]$ using 100 mesh points and CFL= 0.9. As boundary conditions, the stationary solution is imposed at ghost cells.

	ACAT2		WBACAT2
Points	Order	Error	Error
100	-	3.16E-3	7.69E-16
200	2.02	7.78E-4	1.81E-15
400	2.01	1.93E-4	1.07E-15
800	2.00	4.81E-5	1.59E-15
1600	2.00	1.19E-5	2.43E-15

Table 4.3: Test 4.5.2 (Preservation of a stationary solution with linear H). Errors in L^1-norm and convergence rates at time $t = 8$: ACAT2 and WBACAT2.

Figure 4.5.4 shows the differences between the exact and the numerical solutions obtained with ACAT2P, WBACAT2P, $p = 1, 2$ at time $t = 8$. While the non well-balanced schemes give accurate solutions according to their order, the well-balanced methods capture the

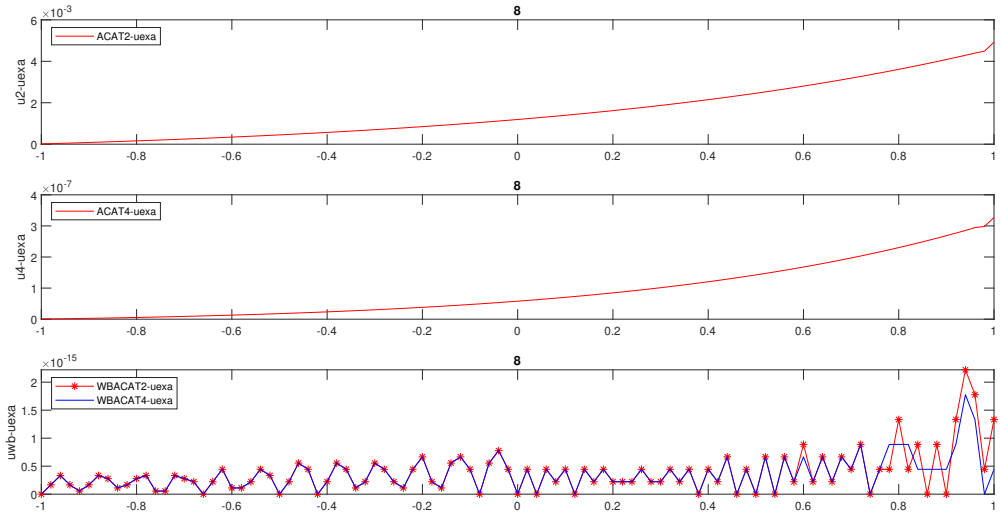

Figure 4.5.4: Test 4.5.2 (Preservation of a stationary solution with linear H). Differences between the exact and the numerical solutions at time $t = 8$ using a mesh of 100 points and CFL= 0.9 : ACAT2 (top), ACAT4 (center), WBACAT2 and WBACAT4 (bottom).

	ACAT4		WBACAT4
Points	Order	Error	Error
100	-	1.81E-7	7.19E-16
200	4.02	1.12E-8	2.16E-15
400	4.01	6.91E-10	3.68E-15
800	4.00	4.31E-11	5.59E-15
1600	3.99	2.71E-12	3.45E-15

Table 4.4: Test 4.5.2 (Preservation of a stationary solution with linear H). Errors in L^1-norm and convergence rates at time $t = 8$: ACAT4 and WBACAT4.

stationary solution with machine precision. This behaviour is also shown on Tables 4.3-4.4 where the L^1-errors and the empirical order of convergences, corresponding to ACAT2P and WBACAT2P with $P = 1, 2$, at time $t = 8$ are exhibited, emphasizing the good properties of the well-balanced schemes.

Perturbation of a stationary solution with linear H

Let us consider equation (4.5.6) with $H(x) = x$ and initial condition a small perturbation of the stationary solution (see Figure 4.5.5):

$$u_0(x) = e^x + 0.0002e^{-200(x+0.7)^2} \tag{4.5.7}$$

The problem is solved in the interval $[-1, 1]$ at time $t = 1$ using 100 mesh points and

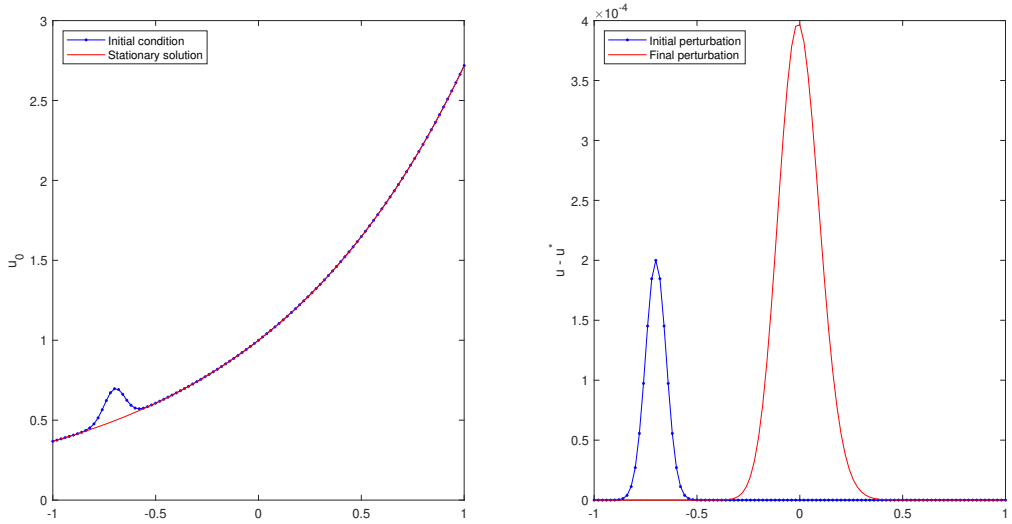

Figure 4.5.5: Test 4.5.2 (Perturbation of a stationary solution with linear H). Initial condition and stationary solution (left). Differences between numerical solution and stationary one at initial and final time (right). The perturbation of the initial condition (left) is amplified by 1000 times in order to see clearly the perturbation. The reference solution is computed with WBACAT4 adopting a 2000 mesh points and CFL= 0.9 at time $t = 1$.

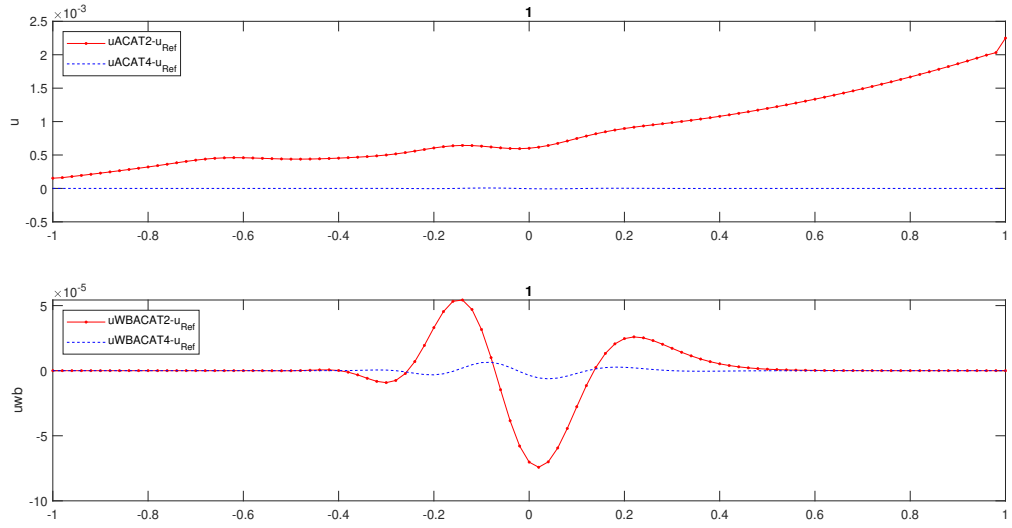

Figure 4.5.6: Test 4.5.2 (Perturbation of a stationary solution with linear H). Differences between the reference solution obtained by WBACAT4 using a 2000 mesh points and the numerical solutions computed at time $t = 1$ using 100 mesh points and CFL $= 0.9$: ACAT2-4 (top) and WBACAT2-4 (bottom).

CFL= 0.9. As boundary conditions, the stationary solution is imposed at ghost cells.

As it can be seen, Figure 4.5.6 shows the difference between a reference solution computed with WBACAT4 adopting a 2000 mesh point and the numerical solutions provided by ACAT2P and WBACAT2P, with $P = 1, 2$. All the schemes are able to evolve the initial perturbation in accordance with their order. In particular, it can be observed that the well-

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
25	-	2.76E-2	-	8.95E-5	-	8.91E-4	-	6.91E-5
50	1.99	6.93E-3	0.79	5.18E-5	2.17	1.98E-4	1.90	1.85E-5
100	2.00	1.73E-3	1.31	2.09E-5	3.45	1.80E-5	3.43	1.71E-6
200	2.00	4.31E-4	1.76	6.15E-6	3.90	1.21E-6	3.89	1.15E-7
400	2.01	1.07E-4	1.93	1.61E-6	3.99	7.62E-8	3.98	7.29E-9

Table 4.5: Test 4.5.2: (Perturbation of a stationary solution with linear H). Errors in L^1-norm and empirical convergence rates for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 1$ and CFL$= 0.9$.

balanced solutions have an smaller error than the non well-balanced version. This behaviour, and the computation of empirical convergence rates, could be also checked in Table 4.5.

In this experiment we have checked the well-balanced property and the order accuracy choosing H linear. Our new question is what happen if we consider a non-linear expression for H?

Preservation of a stationary solution with non-linear H

To answer to the previous question, let us consider the Burgers equation (4.5.6) with a non-linear H

$$H(x) = x + 0.1\sin(10x).$$

Figure 4.5.7 shows the differences between the exact and the numerical solutions obtained with ACAT2P, WBACAT2P, $P = 1, 2$ at time $t = 8$. While the non well-balanced schemes give accurate solutions according to their order, the well-balanced methods capture the stationary solution with machine precision.

The first step is to check the well-balanced property. Indeed, we solve (4.5.6) with initial condition the stationary solution

$$u_0(x) = e^{H(x)}$$

in the interval $[-1, 1]$ using 100 mesh points and CFL$= 0.9$. As boundary conditions, the stationary solution is imposed at ghost cells.

Table 4.6 shows the L^1-errors and the empirical order of convergences corresponding to ACAT2P, and the errors of WBACAT2P methods, with $P = 1, 2$ at time $t = 8$. While the non well-balanced schemes give accurate solutions according to their order, the well-balanced methods capture the stationary solution with machine precision.

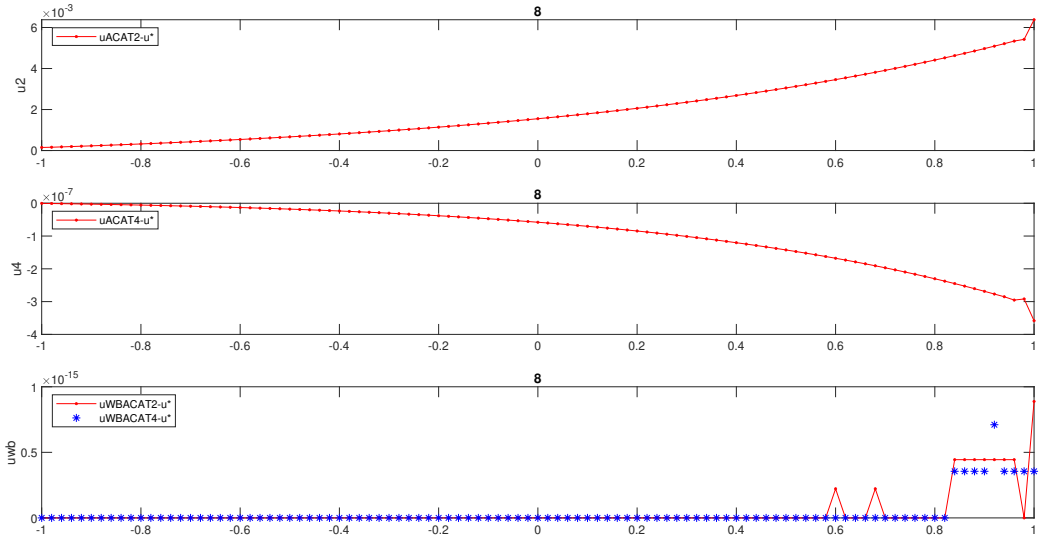

Figure 4.5.7: Test 4.5.2. Differences between the exact and the numerical solutions at time $t = 8$ using a mesh of 100 points and CFL= 0.9: ACAT2 (top), ACAT4 (center), WBACAT2 and WBACAT4 (bottom).

	ACAT2		WBACAT2	ACAT4		WBACAT4
Points	Order	Error	Error	Order	Error	Error
100	-	1.82E-3	2.66E-17	-	1.93E-5	3.99E-17
200	2.05	4.32E-4	3.11E-17	4.09	1.13E-6	6.21E-17
400	2.03	1.07E-4	2.44E-17	4.05	6.84E-8	2.22E-17
800	2.01	2.63E-5	2.78E-17	4.03	4.19E-9	3.55E-17
1600	2.00	6.55E-6	1,99E-17	4.01	2.59E-10	5.21E-17

Table 4.6: Test 4.5.2. (Preservation of a stationary solution with non-linear H). Errors in L^1-norm and convergence rates at time $t = 8$ for ACAT2-4. The errors for WBACAT2-4 are due to round-off.

Checked that the well-balanced property for the Burgers equation with non-linear H is satisfied, we will now focus on some experiment to test the accuracy of the numerical solutions where different types of initial conditions are considered.

Perturbation of a stationary solution with non-linear H

Let us consider (4.5.6) with oscillatory H given by

$$H(x) = x + 0.1\sin(10x)$$

and initial condition

$$u_0(x) = e^{H(x)} + 0.02e^{-200(x+0.7)^2}, \tag{4.5.8}$$

that is a small smooth perturbation of the stationary solution $u^*(x) = e^{H(x)}$: see Figure 4.5.8. We solve the problem in the interval $[-1, 1]$ using 200 mesh points and CFL= 0.9. As

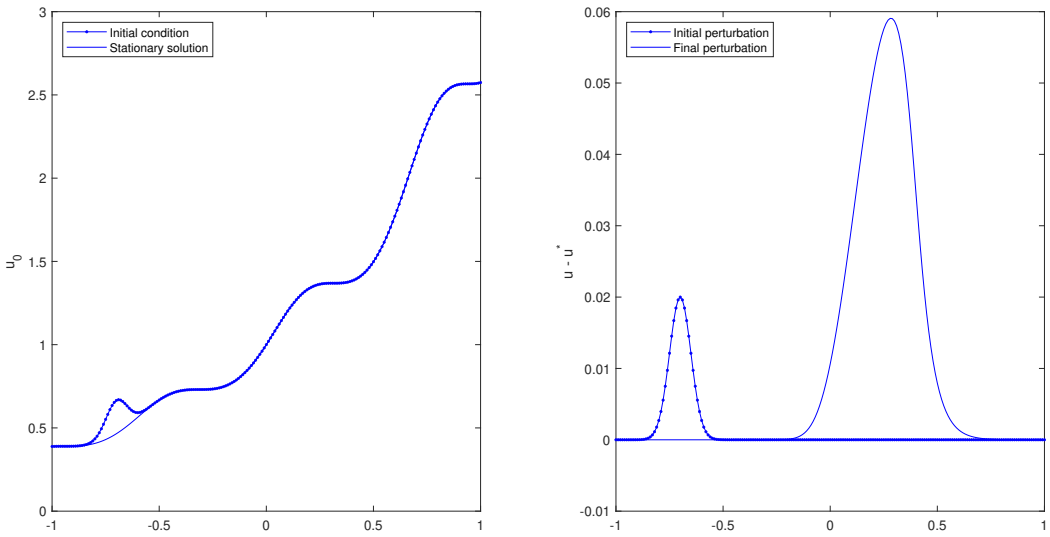

Figure 4.5.8: Test 4.5.2. (Perturbation of a stationary solution with non-linear H). Initial condition and stationary solution (left). Differences between reference and stationary solution at initial and final time (right). The perturbation of the initial condition (left) is amplified by 10 times in order to see clearly the perturbation. The reference solution is computed with WBACAT4 using 1000 mesh points and CFL= 0.9 at time $t = 1.2$.

boundary conditions the stationary solution is imposed at left ghost point and free boundary at right.

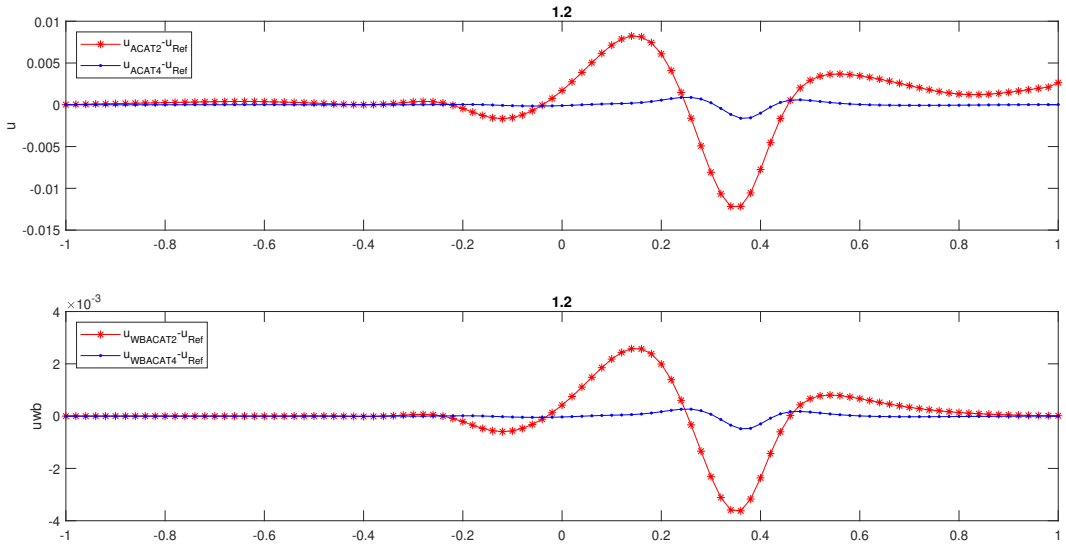

Figure 4.5.9: Test 4.5.2. (Perturbation of a stationary solution with non-linear H). Differences between numerical solutions computed with ACAT2P (top) and WBCAT2P, (bottom) $P = 1, 2$, and reference solution at $t = 1.2$ using a mesh of 200 points and CFL= 0.9. For the reference solution the WBACAT4 is adopted with a mesh of 1000 points.

Figure 4.5.9 shows that, all the schemes are able to evolve the perturbation in according with the order. Nevertheless, the well-balanced methods, WBACAT2 and WBACAT4, are able to capture more precisely the evolution of the perturbation with a smaller error than the relative non well-balanced schemes.
For the errors in L^1-norm and convergence rates, we adopt as initial condition

$$u_0(x) = e^{H(x)} + 0.2e^{-200(x+0.7)^2},$$

in other word, a bigger perturbation is considered reducing the final time to $t = 0.2$.
Table 4.7 shows that the non well-balanced methods introduce a bigger error in comparison with the well-balanced approach but an increasing of points is necessary to achieve the theoretical order. This phenomenon is partly attributable to reconstruction partly to smoothness indicators, because they fail to detect the theoretical regularity.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
80	-	6.44E-2	-	1.07E-2	-	1.53E-3	-	1.17E-4
160	1.15	2.88E-2	1.74	3.21E-3	2.11	3.10E-4	2.03	2.86E-5
320	1.45	1.05E-2	1.89	8.61E-4	4.28	1.59E-5	4.45	1.31E-6
640	1.62	3.42E-3	1.96	2.21E-4	4.14	8.99E-7	3.85	8.99E-8
1280	1.86	9.41E-4	1.99	5.58E-5	4.01	5.57E-8	3.96	5.76E-9

Table 4.7: Test 4.5.2: (Perturbation of a stationary solution with non-linear H). Errors in L^1-norm and convergence rates for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.2$ and CFL= 0.9.

This experiment is very interesting because, on the one hand it shows that the well-balanced methods work well near the stationary solution even in case H is non-linear, on the other hand, it happens that a small perturbation of the initial state, although rather smooth, may result in a loss of accuracy in the adaptive order reconstruction when a not fine mesh grid is adopted. To emphasize this behaviour, a non-linear oscillatory H is considered in next experiments.

Preservation of a stationary solution with oscillatory H

Following [97], we consider (4.5.6) with

$$H(x) = x + \frac{1}{10}\sin(100x), \tag{4.5.9}$$

and we take as initial condition the stationary solution

$$u^*(x) = e^{H(x)}$$

(see Figure 4.5.10). We solve the problem in the interval $[-1, 1]$ using 100 mesh points and

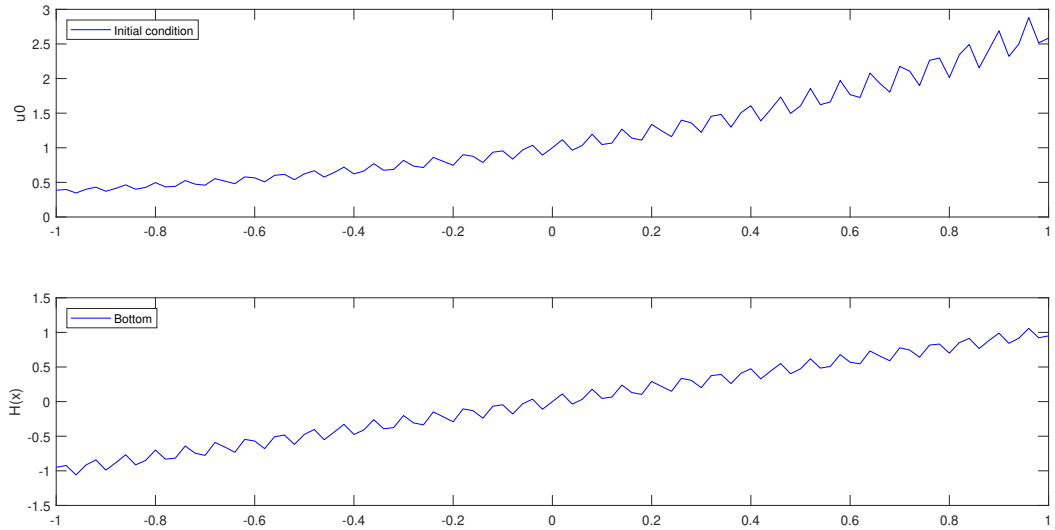

Figure 4.5.10: Test 4.5.2. (Preservation of a stationary solution with oscillatory H). Initial condition (top) and H (down).

CFL= 0.9. With this choice of mesh points, the period of the oscillations of H is close to Δx. As boundary conditions the stationary solution is imposed again at ghost points.

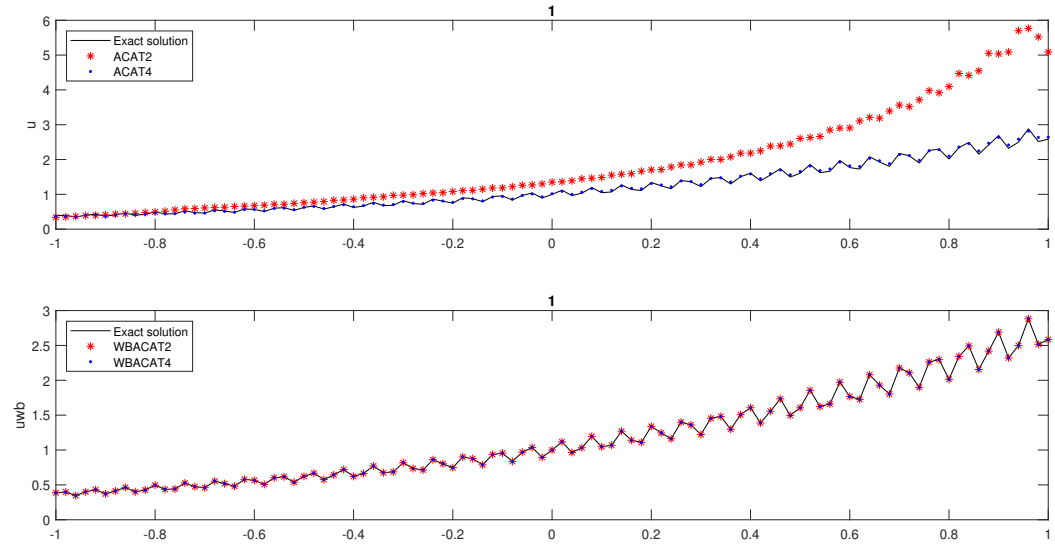

Figure 4.5.11: Test 4.5.2. (Preservation of a stationary solution with oscillatory H). Exact and numerical stationary solutions computed with ACAT2P and WBCAT2P at $t = 1$ using a mesh of 100 points and CFL= 0.9: non well-ballanced (top) and well-balanced (bottom).

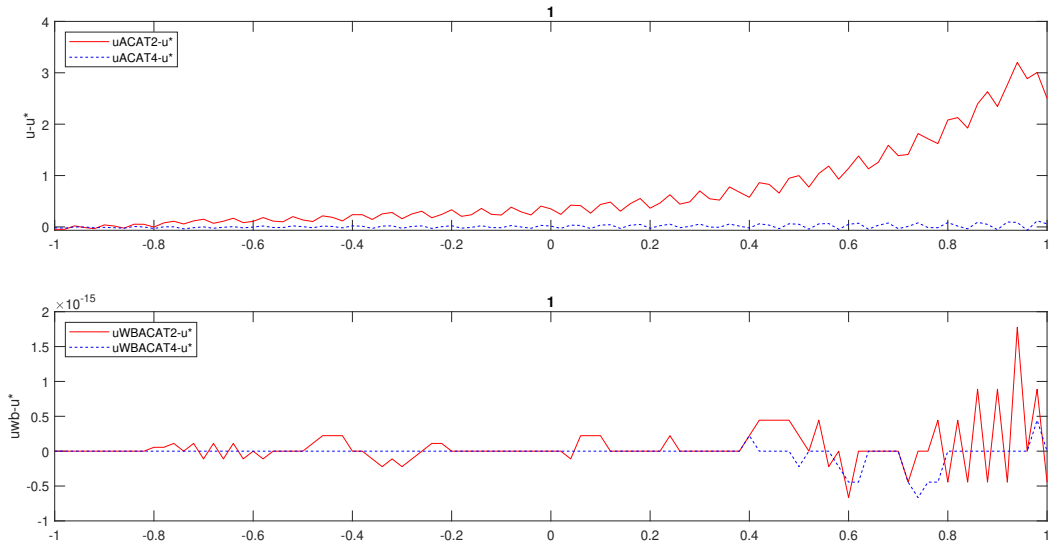

Figure 4.5.12: Test 4.5.2. (Preservation of a stationary solution with oscillatory H). Top: differences between the exact and the numerical stationary solutions computed with ACAT2P, $P = 1, 2$, at time $t = 1$ using 100 mesh points and $CFL = 0.9$. Bottom: differences between the exact and the numerical solutions computed with WBACAT2P, $P = 1, 2$, at time $t = 1$ using 100 mesh points and CFL= 0.9.

Figure 4.5.11 shows that, while WBACAT2 and WBACAT4 capture the stationary solution with machine precision, this is not the case for ACAT2 and ACAT4. Figure 4.5.12 displays the differences between the numerical solutions and the stationary solution obtained at time $t = 1$ (top). In this case, the results provided by WBACAT2 and WBACAT4 (bottom) are very similar and they are able to capture the machine precision; while, the results provided by ACAT2 and ACAT4 (top) show that the non well-balanced methods are not able to detect the stationary solution with high precision even if a 4 order method is applied.

Perturbation of a stationary solution with oscillatory H

We consider again (4.5.6) with oscillatory H given by (4.5.9) (see Figure 4.5.10) and initial condition a small perturbation of the stationary solution

$$u_0(x) = e^{H(x)} + 0.0002e^{-200(x+0.7)^2}. \tag{4.5.10}$$

We solve the problem in the interval $[-1, 1]$ at time $t = 1.25$ adopting a 200 mesh points and CFL= 0.9. As boundary conditions the stationary solution is imposed at ghost points. Figure 4.5.13 shows the stationary solution and the perturbation of the stationary solution, amplified by 1000 times, (left); the initial and the final signal (right). The reference solution

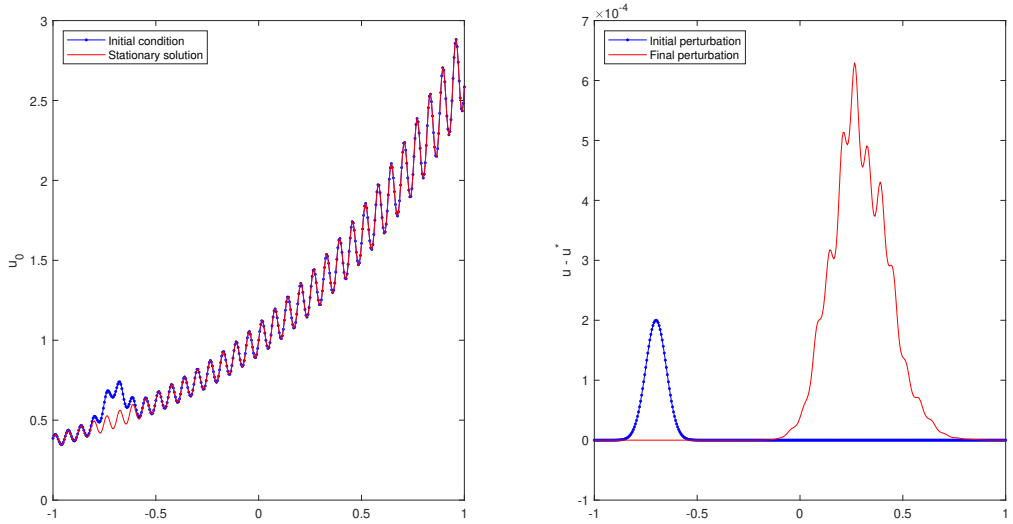

Figure 4.5.13: Test 4.5.2 (Perturbation of a stationary solution with oscillatory H). Initial condition and stationary solution (left). Differences between reference and stationary solution at initial and final time (right). The perturbation of the initial condition (left) is amplified by 1000 times in order to see clearly the perturbation. The reference solution is computed with WBACAT4 using 8000 mesh points and CFL= 0.9 at time $t = 1.25$.

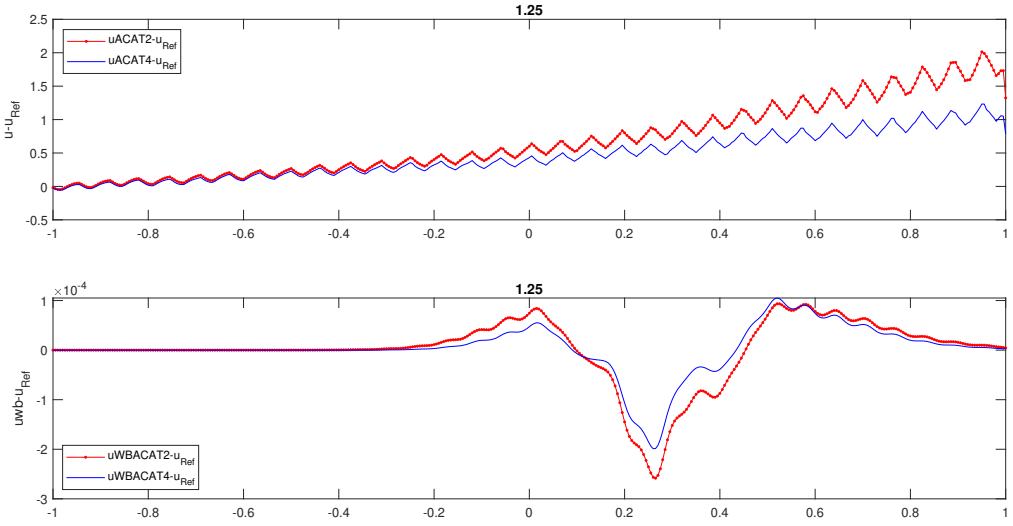

Figure 4.5.14: Test 4.5.2 (Perturbation of a stationary solution with oscillatory H). Differences between the reference and numerical solutions computed with ACAT2P (top) and WBCAT2P (bottom), $P = 1, 2$, at $t = 1.25$ using a 200 mesh points and CFL= 0.9. The reference solution is computed with WBACAT4 using 8000 mesh points and CFL= 0.9 at time $t = 1.25$.

has been computed with WBACAT4 using a 8000 mesh points and CFL= 0.9.

Figure 4.5.14 shows the difference between the numerical solution computed with ACAT2P (top) and WBACAT2P (bottom), with $P = 1, 2$, adopting a 200 mesh points and the reference solution. As it can be observed, the well-balanced schemes are able to capture the correct solution with a lower error than the non well-balanced one. Nevertheless, there is

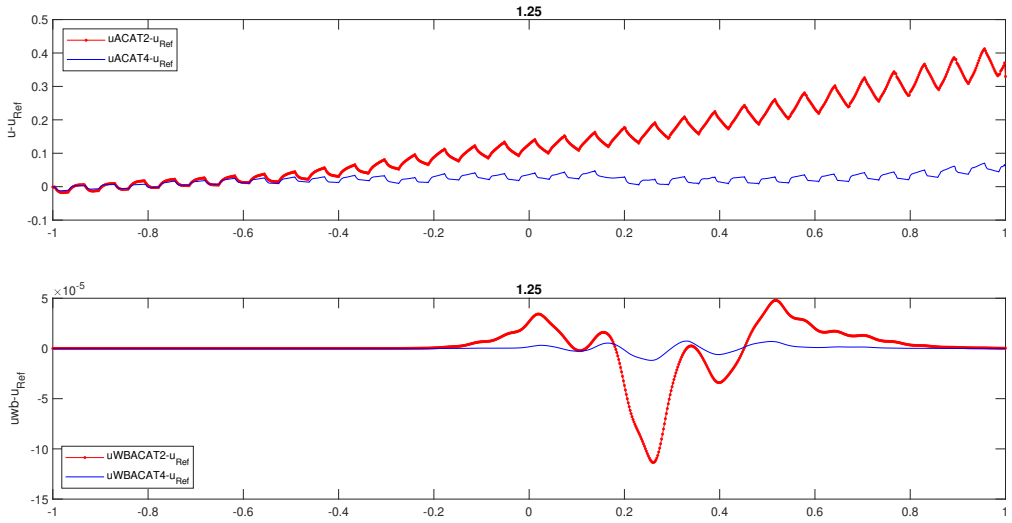

Figure 4.5.15: Test 4.5.2 (Perturbation of a stationary solution with oscillatory H). Differences between the reference and numerical solutions computed with ACAT2P (top) and WBCAT2P (bottom), $P = 1, 2$, at $t = 1.25$ using a 800 mesh points and CFL$= 0.9$. The reference solution is computed with WBACAT4 using 8000 mesh points and CFL$= 0.9$ at time $t = 1.25$.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
125	-	1.89E0	-	1.88E-4	-	1.89E0	-	1.88E-4
250	-1.47	5.24E0	0.74	1.12E-4	-1.34	4.78E0	0.74	1.12E-4
500	2.40	9.95E-1	0.99	5.65E-5	3.55	4.08E-1	2.07	2.67E-5
1000	1.85	2.76E-1	1.33	2.25E-5	3.12	4.69E-2	3.31	2.68E-6
2000	1.80	7.94E-2	1.17	9.97E-6	5.04	1.43E-3	9.29	4.26E-9
4000	1.84	2.22E-2	1.41	3.76E-6	15.09	4.08E-8	5.73	8.01E-11

Table 4.8: Test 4.5.2: (Perturbation of a stationary solution with oscillatory H). Errors in L^1-norm and numerical convergence rates for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 1.25$ and CFL$= 0.9$.

almost no difference between the solution obtained with WBACAT2 and WBACAT4. This behaviour is due to the smoothness indicators because, with a low fine mesh, they are not able to capture the theoretical smoothness and, as consequence, there is an order reduction in the adaptive strategy. This attitude is also reproduced by the non well-balanced scheme. As it can be seen in Figure 4.5.15, the order reduction phenomenon should be circumvent adopting a finer mesh points that is also confirmed by Table 4.8. In particular, crossing between the 1000 mesh points and the 2000 one, an important decreasing in the error is observed because, from now on, the smoothness indicators capture, as expected, the regularity of the numerical data.

As final comment is remarkable that, at least, all the method should converge to the expected

theoretical order under the hypothesis that the numerical data are smooth enough and in accordance with the smoothness indicators. Nonetheless, the non-well balanced schemes introduce a bigger error compared with the mesh size making, in this cases, the use of well-balanced methods are extremely necessary.

As a final check we consider the behaviour of the methods in the case of an initial condition of class $\mathcal{C}^5$ which is far from the stationary solution.

Burgers Order Test

Let us consider (4.5.6) with $H(x) = x$ and initial condition (3.2.24) (see Figure 4.5.16)

$$u_0(x) = \begin{cases} 0 & \text{if} \quad x < 0; \\ p(x) & \text{if} \quad 0 \leq x \leq 1; \\ 1 & \text{if} \quad x > 1; \end{cases} \tag{4.5.11}$$

where

$$p(x) = x^6 \Big(\sum_{k=0}^{5} (-1)^k \binom{5+k}{k} (x-1)^k \Big).$$

We solve the problem in the interval $[-0.2, 2]$ using 80 mesh points and CFL= 0.9. As boundary conditions free boundary is imposed at ghost points.

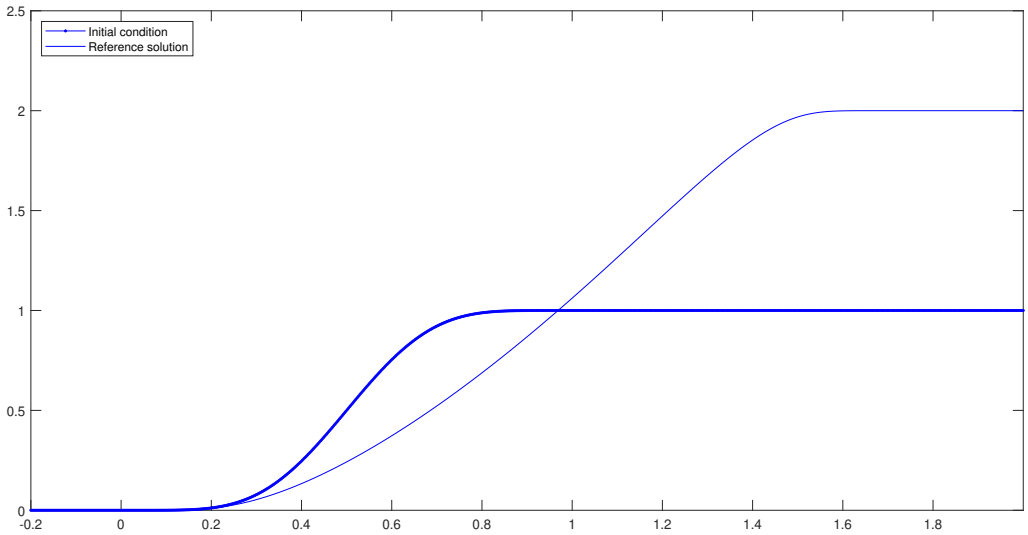

Figure 4.5.16: Test 4.5.2 (Burgers Order Test). Initial condition and reference solution obtained with WBACAT4 using a 2560 mesh points and CFL= 0.9 at time $t = 0.5$.

Figure 4.5.17 and Table 4.9 show us how all the methods manage to produce solutions in agreement with each other obtaining the expected order.

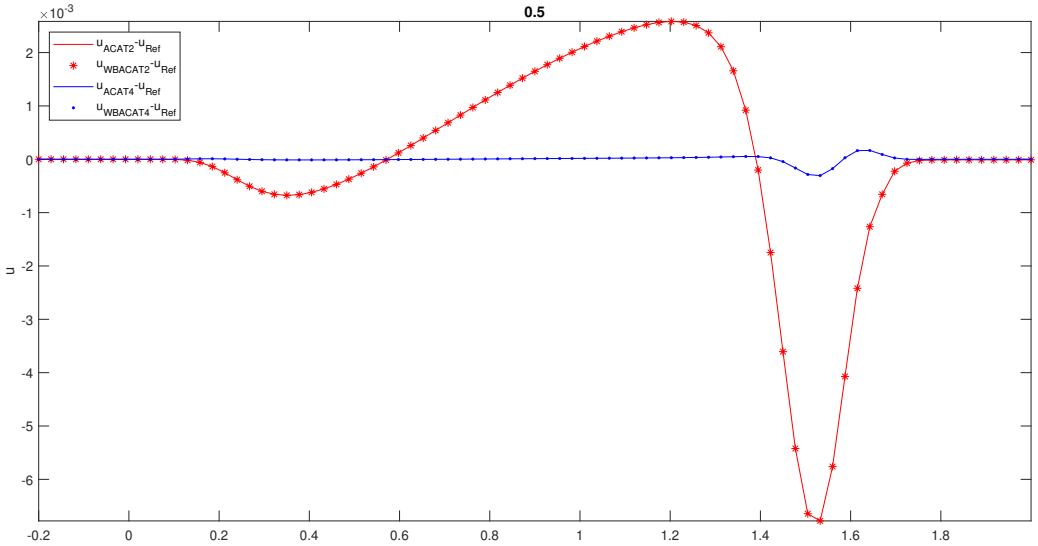

Figure 4.5.17: Test 4.5.2 (Burgers Order Test). Differences between numerical solutions computed with ACAT2P and WBCAT2P, $P = 1, 2$, and the reference solution at $t = 0.5$ using a mesh of 80 points and CFL= 0.9. For the reference solution a mesh of 2560 points has been adopted.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
80	-	1.74E-3	-	1.91E-3	-	2.41E-4	-	2.42E-4
160	1.93	4.11E-4	1.92	5.05E-4	2.40	4.56E-5	2.41	4.57E-5
320	1.93	1.08E-4	1.94	1.32E-4	3.17	5.05E-6	3.18	5.06E-6
640	1.98	2.74E-5	1.98	3.34E-5	3.68	3.95E-7	3.68	3.94E-7
1280	1.99	6.91E-6	1.99	8.41E-6	3.91	2.60E-8	3.92	2.61E-8
2560	2.00	1.73E-6	2.00	2-11E-6	3.98	1.65E-9	3.98	1.65E-9

Table 4.9: Test 4.5.2: (Burgers Order Test) Errors in L^1-norm and convergence rates for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.5$ and CFL= 0.9.

This experiment shows that, when the initial condition is far from the stationary solution, well-balanced and non well-balanced methods produce essentially the same results, with the expected order of accuracy.

4.5.3 Shallow water model

In this section we will focus on the one-dimensional hyperbolic shallow water model

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\right)_x = ghH_x, \end{cases} \tag{4.5.12}$$

that can be written in the form (4.0.1) with

$$U = \begin{bmatrix} h \\ q \end{bmatrix}, \quad f(U) = \begin{bmatrix} q \\ \dfrac{q^2}{h} + \dfrac{g}{2}h^2 \end{bmatrix}, \quad S(U) = \begin{bmatrix} 0 \\ gh \end{bmatrix}.$$

The variable x makes reference to the axis of the channel and t is time; $q(x,t)$ and $h(x,t)$ represent the mass-flow and the thickness; g, the acceleration due to gravity; $H(x)$, the depth measured from a fixed level of reference; furthermore, the following relation is verified $q(x,t) = h(x,t)u(x,t)$, with u the depth average horizontal velocity. The eigenvalues of the Jacobian matrix $J(U)$ of the flux $f(U)$ are

$$\lambda_1 = u - \sqrt{gh} \quad \text{and} \quad \lambda_2 = u + \sqrt{gh}.$$

The Froude number is defined by

$$Fr = \frac{|u|}{\sqrt{gh}}.$$

The flow is said to be supercritical if $Fr > 1$, critical if $Fr = 1$, and subcritical if $Fr < 1$.

The stationary solution of the shallow water system (4.5.12) are implicitly given by

$$q = C_1 \quad \text{and} \quad \frac{1}{2}\frac{q^2}{h^2} + gh - gH = C_2, \tag{4.5.13}$$

where C_1 and C_2 are arbitrary constants [17, 20]. In order to implement the well-balanced methods, given $U_i^n = [h_i^n, q_i^n]^T$ one has to find the stationary solution $U_i^* = [q_i^*, h_i^*(x)]^T$ that solves (4.3.6): it is implicitly given by

$$q_i^*(x) = q_i^n, \quad \frac{1}{2}\frac{(q_i^n)^2}{h_i^*(x)^2} + gh_i^*(x) - gH = C_i, \quad \forall x,$$

with

$$C_i = \frac{1}{2}\frac{(q_i^n)^2}{(h_i^n)^2} + gh_i^n - gH(x_i).$$

Therefore, at a point x_j of the stencil, one has $q_i^*(x_j) = q_i^n$ and $h_i^*(x_j)$ has to be a positive root of the polynomial:

$$P_{i,j}(h) = h^3 - \left(\frac{C_i}{g} + gH(x_j)\right)h^2 + \frac{1}{2g}(q_i^n)^2.$$

This polynomial can have two, one, or zero positive roots. In the first case, one of the roots corresponds to a supercritical state and the other one to a subcritical state: a criterion is necessary to select one root or the other. We follow here a similar criterion to the one chosen in [20] in the context of finite volume methods: the solution whose regime (sub or supercritical) is the same as the one of U_i^n is selected [1]. A careful implementation is needed to capture transcritical stationary solutions: see for instance the discussion in [20] or [97].

Preservation of a subcritical stationary solution

Let us consider the shallow water model in the space interval $[-3, 3]$ with bottom depth given by

$$H(x) = \begin{cases} -0.25(1 + \cos(5\pi x)) & \text{if } -0.2 \leq \mathrm{x} \leq 0.2; \\ 0 \quad \text{otherwise;} \end{cases} \tag{4.5.14}$$

and initial condition given by the subcritical stationary solution U^* that satisfies

$$q^* = 2.5, \quad h^*(-3) = 2$$

(see Figure 4.5.18). The numerical methods are applied to this problem using 200 mesh points and CFL= 0.8. At the boundaries, the stationary solution is imposed at ghost points.

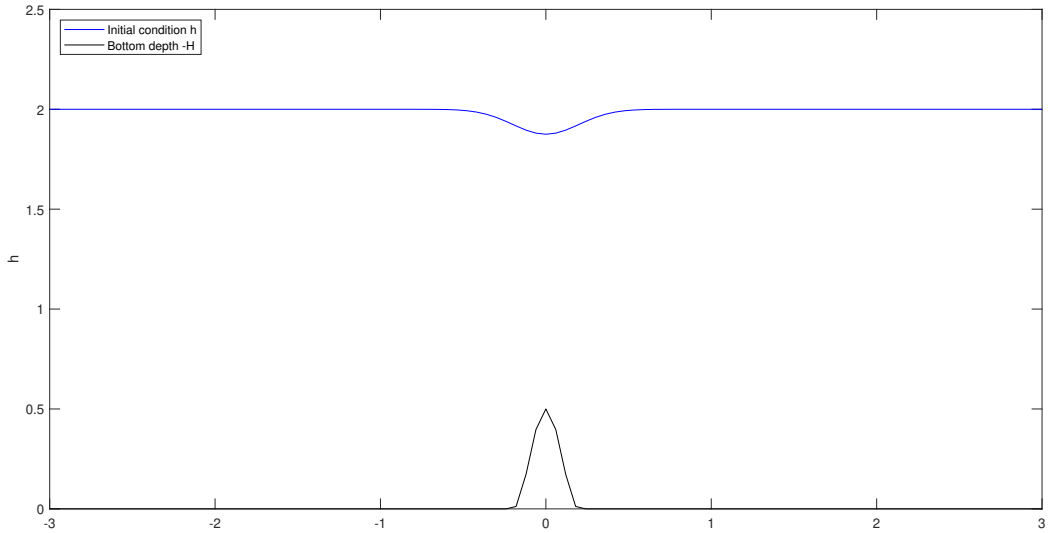

Figure 4.5.18: Test 4.5.3. (Preservation of a subcritical stationary solution). Discrete initial condition with 100 mesh points. Free surface and bathymetry.

Figure 4.5.20 shows the differences between the exact and the non well-balanced numerical solutions obtained at time $t = 4$. As it can be seen in Figure 4.5.19, the well-balanced

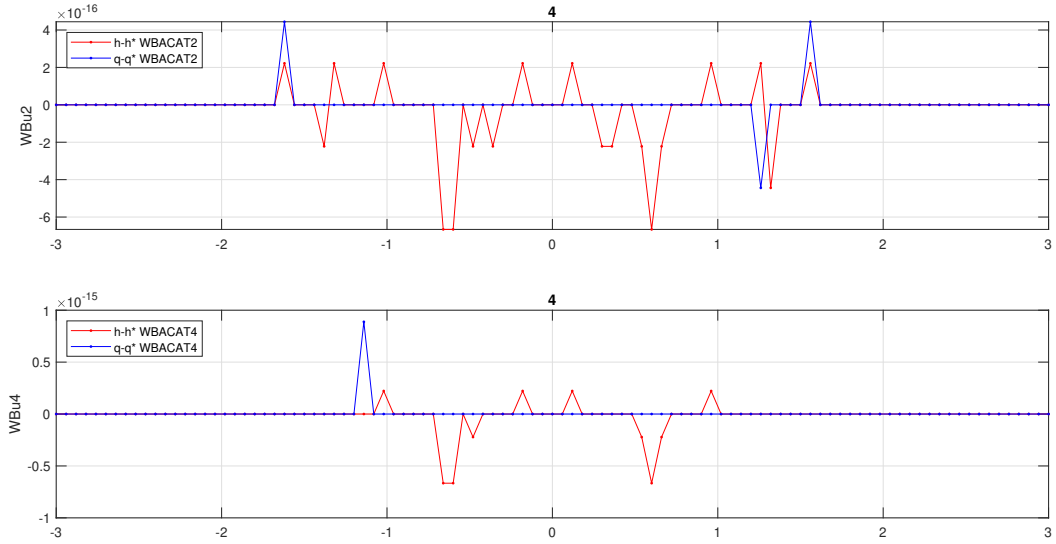

Figure 4.5.19: Test 4.5.3. (Preservation of a subcritical stationary solution). Differences between the numerical solutions for second order (top) and fourth order (bottom) obtained with well-balanced, WBACAT, methods and the exact stationary one, at time $t = 4$ using 100 mesh points and CFL= 0.8.

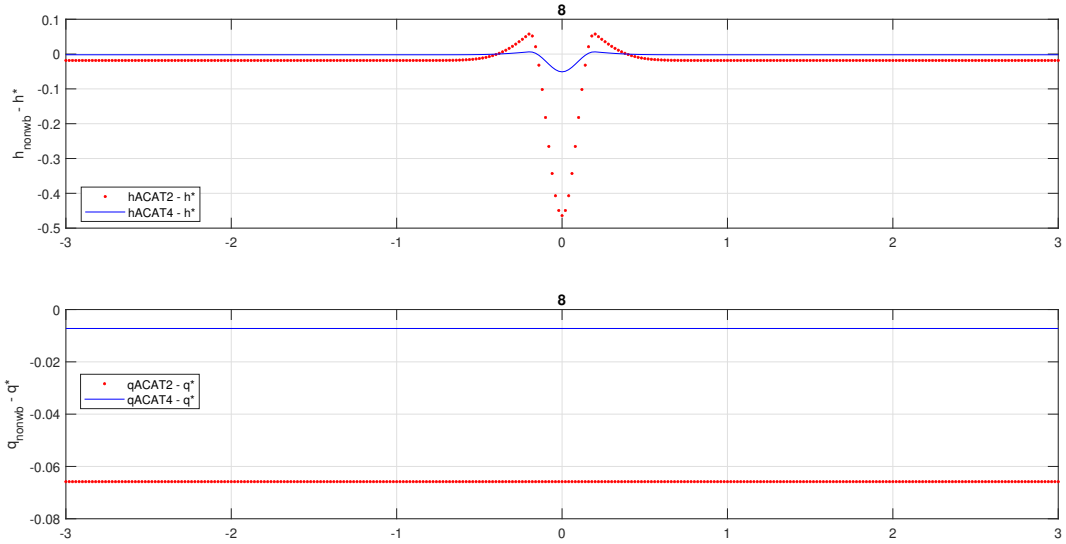

Figure 4.5.20: Test 4.5.3. (Preservation of a subcritical stationary solution). Differences between the numerical solutions for h (top) and q (bottom) obtained with ACAT methods and the exact stationary one, at time $t = 4$ using 200 mesh points and CFL= 0.8.

methods capture the stationary solution to machine accuracy even with a not fine mesh grid. This behaviour is confirmed by Table 4.10 that shows the L^1-errors corresponding to WBACAT2P, $P = 1, 2$, using 50, 100, 200 and 400 mesh points at time $t = 4$. Unfortunately, the introduction of spurious oscillations with the not well-balanced schemes involves a order reduction since the high order smoothness indicators are not able to detect a priori the real smoothness of the solution. This behaviour is highlighted in the next experiments.

	WBACAT2			WBACAT4		
Points	h	q	u	h	q	u
50	2.93E-16	1.07E-16	2.66E-16	2.39E-16	5.32E-17	1.87E-16
100	3.46E-16	7.99E-17	1.86E-16	2.13E-16	0	1.20E-16
200	3.40E-16	0	2.46E-16	3.99E-17	0	1.99E-17
400	1.77E-16	0	1.20E-16	0	5.99E-17	2.98E-17

Table 4.10: Test 4.5.3. (Preservation of a subcritical stationary solution). Errors in L^1-norm for WBACAT2P, $P = 1, 2$, at time $t = 4$.

Perturbation of a subcritical stationary solution

The setting of this test is similar to the previous one but now the initial condition is a smooth perturbation of the subcritical stationary solution U^* (see Figure 4.5.21) considered there:

$$U_0 = \begin{bmatrix} h^* + 0.006e^{(-20(x+1)^2)} \\ q^* \end{bmatrix}.$$

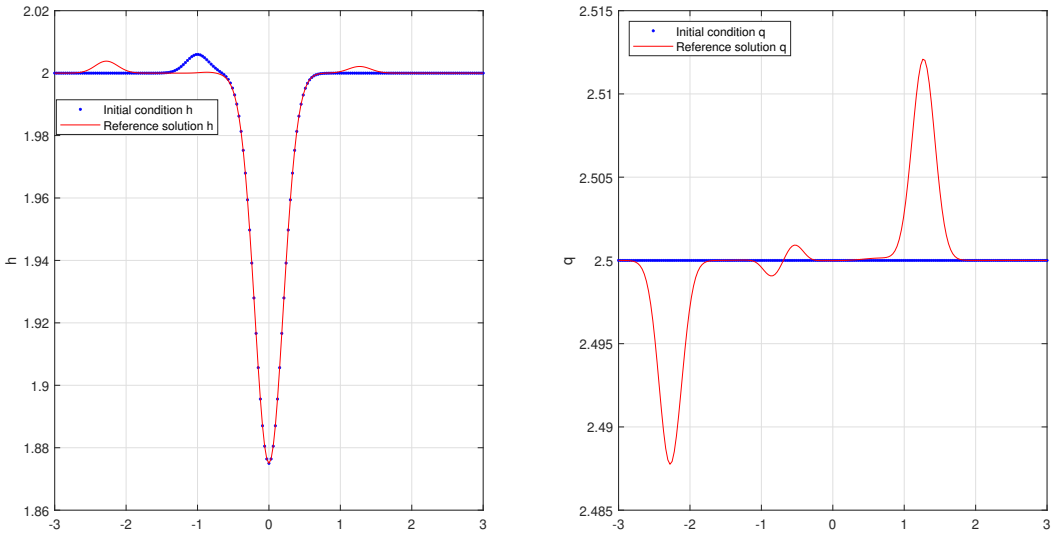

Figure 4.5.21: Test 4.5.3. (Perturbation of a subcritical stationary solution). Initial condition and reference solution obtained with WBACAT4 computed at time $t = 0.4$ using 2000 mesh points and CFL$= 0.8$: h (left); q (right). In the plot of q there appear the left and right traveling waves, as well as a small left moving reflected wave.

The numerical solutions are computed on the interval $[-3, 3]$ using 200 mesh points at time $t = 0.4$ with CFL $= 0.8$. As boundary conditions the subcritical stationary solution is imposed at ghost points.

Figures 4.5.22 and 4.5.23 show the errors obtained by the differences between reference

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
50	-	2.18E-3	-	1.64E-5	-	2.07E-4	-	1.81E-5
100	0.98	1.10E-3	1.95	4.22E-6	1.11	9.61E-5	2.25	3.79E-6
200	1.23	4.96E-4	1.94	1.09E-6	1.59	3.19E-5	4.72	1.44E-7
400	1.48	1.68E-4	1.97	2.77E-7	1.84	8.87E-6	4.33	8.22E-9
800	1.53	5.82E-5	1.97	7.07E-8	1.93	2.31E-6	4.07	6.54E-10

Table 4.11: Test 4.5.3: (Perturbation of a subcritical stationary solution). Errors in L^1-norm and convergence rates related to h for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.15$ and CFL$= 0.8$.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
50	-	8.21E-3	-	5.54E-5	-	7.85E-4	-	6.15E-5
100	1.00	4.11E-3	1.63	1.79E-5	1.22	3.35E-4	2.86	1.69E-5
200	1.25	1.73E-3	1.94	4.68E-6	1.67	1.05E-4	4.78	6.17E-7
400	1.51	6.06E-4	1.98	1.19E-6	1.87	2.86E-5	4.31	4.65E-8
800	1.54	2.08E-4	1.98	3.01E-7	1.98	7.31E-6	4.05	1-41E-9

Table 4.12: Test 4.5.3: (Perturbation of a subcritical stationary solution). Errors in L^1-norm and convergence rates related to q for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.15$ and CFL$= 0.8$.

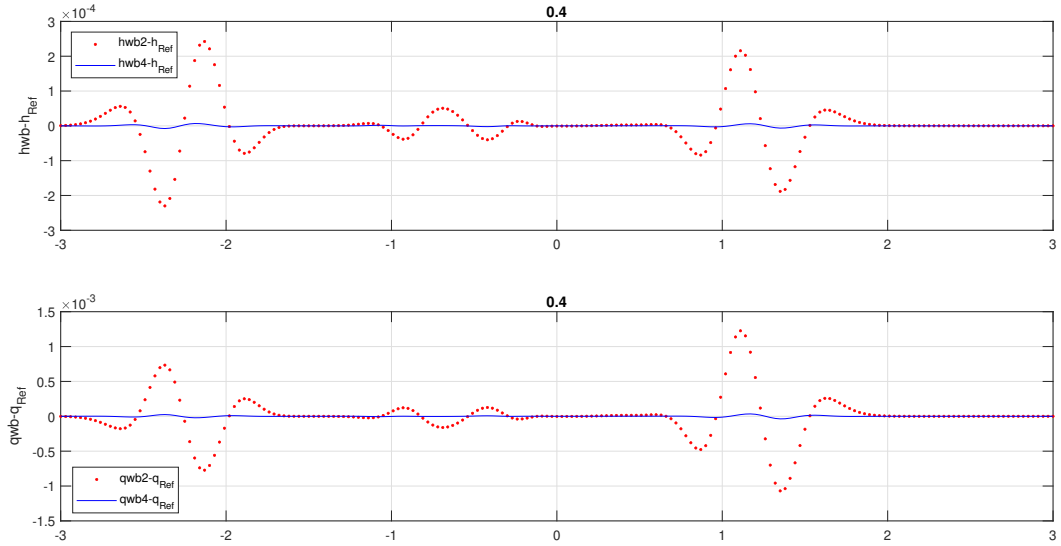

Figure 4.5.22: Test 4.5.3. (Perturbation of a subcritical stationary solution). Differences between reference and numerical solutions obtained with WBACAT2P, $P = 1, 2$, computed at time $t = 0.4$ using 200 mesh points and CFL$= 0.8$: h (top); q (bottom). The reference solution is computed with WBACAT4 adopting a 2000 mesh points.

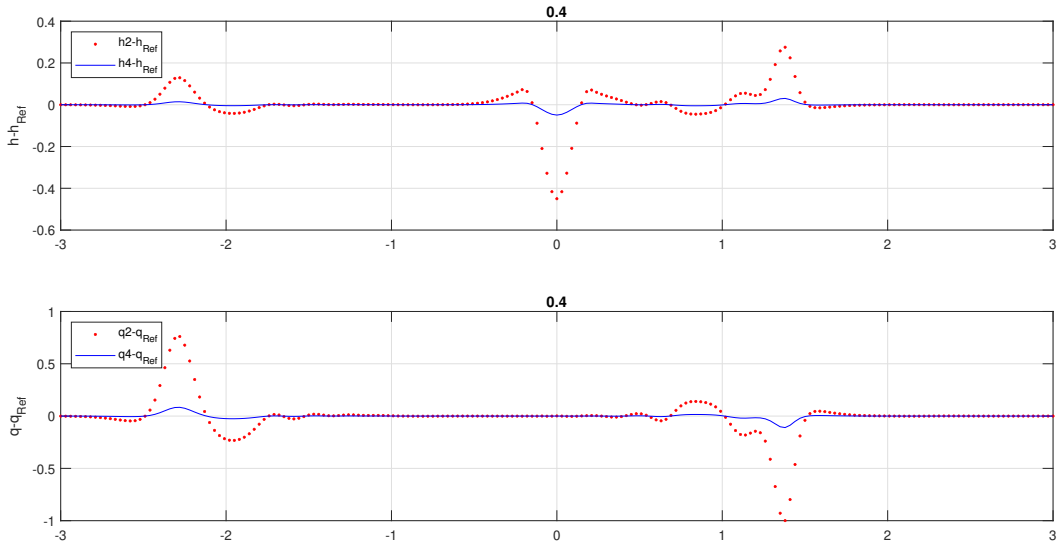

Figure 4.5.23: Test 4.5.3. (Perturbation of a subcritical stationary solution). Differences between reference and numerical solutions obtained with ACAT2P, $P = 1, 2$, computed at time $t = 0.4$ using 200 mesh points and CFL$= 0.8$: h (top); q (bottom). The reference solution is computed with WBACAT4 adopting a 2000 mesh points

solution and the numerical solutions computed with well-balanced and not well-balanced methods at time $t = 0.4$. The reference solution considered is WBACAT4 adopting a 2000 mesh points. As expected, WBACAT2P, $P = 1, 2$, capture better the waves generated by the initial perturbation than ACAT2P, $P = 1, 2$. In addition, Tables 4.11-4.12 show how well-balanced methods manage to reach the expected order, behavior not respected by non well-balanced methods. In this case, as seen above, this phenomenon is partly attributable

to not well-balanced reconstruction partly to smoothness indicators. In fact, the not well-balanced methods at first step introduce a spurious error which implies a loss of numerical smoothness resulting in degradation of the order.

Smooth initial condition with flat bottom

We now check that, in the case of flat bottom and smooth solution, well-balanced and non well-balanced schemes give the same result, all with the expected order of accuracy. In order to obtain these results we consider the Shallow water equation (4.5.12) with flat bottom and smooth initial condition (3.2.24) (see Figure 4.5.24):

$$U_0(x) = \begin{bmatrix} h_0(x) \\ q_0(x) \end{bmatrix}, \tag{4.5.15}$$

where

$$h_0(x) = q_0(x) = \begin{cases} 0 & \text{if} \quad x < 0; \\ p(x) & \text{if} \quad 0 \le x \le 1; \\ 1 & \text{if} \quad x > 1; \end{cases}$$

and

$$p(x) = x^6 \Big(\sum_{k=0}^{5} (-1)^k \binom{5+k}{k} (x-1)^k \Big).$$

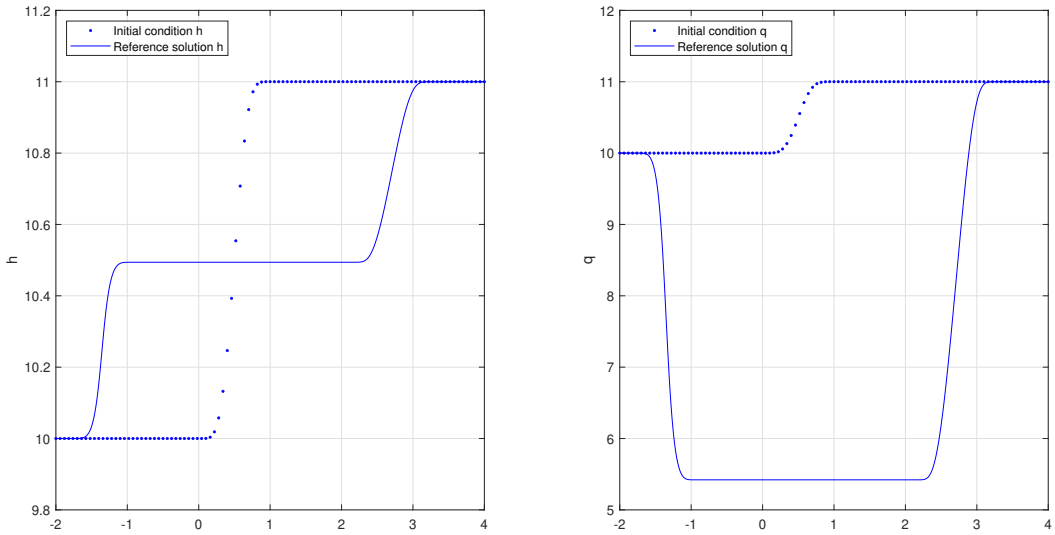

Figure 4.5.24: Test 4.5.3. (Smooth initial condition with flat bottom). Initial condition and reference solution obtained by WBACAT4 at time $t = 0.2$ using 3200 mesh points and CFL= 0.8; h (top); q (bottom).

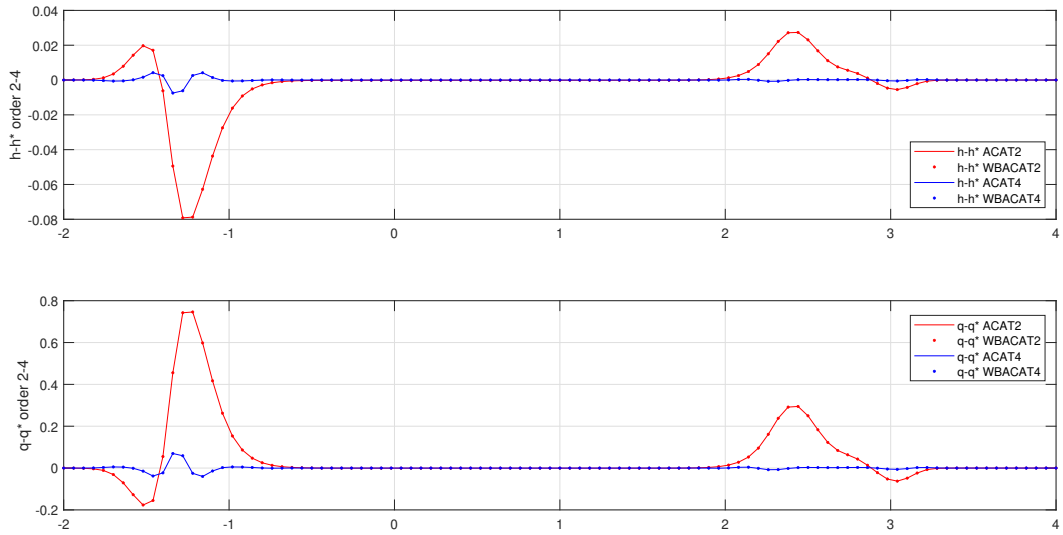

Figure 4.5.25: Test 4.5.3. (Smooth initial condition with flat bottom). Differences between numerical solutions obtained with ACAT2P and WBACAT2P, $P = 1, 2$, computed at time $t = 0.2$ using 100 mesh points and CFL= 0.8 and the reference solution. h (top); q (bottom). For the reference solution a 3200 mesh points has been adopted.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
200	-	7.27E-5	-	7.25E-5	-	2.25E-6	-	2.25E-6
400	2.01	1.80E-5	2.01	1.80E-5	3.98	1.42E-7	3.98	1.42E-7
800	2.00	4.50E-6	2.00	4.50E-6	3.99	8.91E-9	3.99	8,91E-9
1600	2.00	1.13E-6	2.00	1.13E-6	4.00	5.56E-10	4.00	5.57E-10

Table 4.13: Test 4.5.3: (Smooth initial condition with flat bottom). Errors in L^1-norm and convergence rates related to h for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.2$ and CFL= 0.9.

The numerical solutions are computed on the interval $[-2, 4]$ using 100 mesh points and CFL $= 0.9$ at time $t = 0.1$, while for the reference solution WBACAT4 with a 3200 mesh points is adopted. As boundary conditions free boundary is imposed at ghost points.

Figure 4.5.25 and Table 4.13-4.14 show that all methods have a similar behavior by reproducing similar results. In particular, all schemes are accurate as expected.

With this experiment we have proven that all methods, both well-balanced and not, have a similar behavior when they are far from the stationary condition; while, well-balanced reconstructions reproduce better results, both in accuracy and numerical convergence, when we are close to the stationary solution.

	ACAT2		WBACAT2		ACAT4		WBACAT4	
Points	Order	Error	Order	Error	Order	Error	Order	Error
200	-	8.75E-4	-	8.74E-4	-	2.67E-5	-	2.67E-5
400	2.01	2.17E-4	2.01	2.17E-4	3.98	1.69E-6	3.98	1.69E-6
800	2.00	5.42E-5	2.00	5.42E-5	3.99	1.06E-7	3.99	1.06E-7
1600	2.00	1.35E-5	2.00	1.35E-5	4.00	6.61E-9	4.00	6.61E-9

Table 4.14: Test 4.5.3: (Smooth initial condition with flat bottom). Errors in L^1-norm and convergence rates related to q for ACAT2, ACAT4, WBACAT2 and WBACAT4 at time $t = 0.2$ and CFL$= 0.9$.

4.5.4 Euler system with gravity

Let us consider the system of compressible Euler equations of gas dynamics with a gravitational potential in one space dimension

$$\begin{cases} \rho_t + (\rho u)_x = 0 \\ (\rho u)_t + (\rho u^2 + p)_x = -\rho H_x \\ E_t + (u(E+p))_x = -\rho u H_x \end{cases} \tag{4.5.16}$$

Here, ρ is the density, u is the velocity, p is the pressure, E is the energy per unit volume excluding the gravitational energy, and $H(x)$ is the gravitational potential (see [23, 52, 53, 69, 71]). We assume the gas is polytropic, so the pressure is given by

$$p = (\gamma - 1)\Big[E - \frac{1}{2}\rho u^2\Big], \quad \gamma = \frac{c_p}{c_v} > 1,$$

where γ is the ratio of specific heats at constant pressure and volume, which is taken to be constant, in our cases $\gamma = 1.2$. This system can be written in the form (4.0.1) with

$$U = \begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix}, \quad f(U) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{bmatrix}, \quad S(U) = \begin{bmatrix} 0 \\ -\rho \\ -\rho u \end{bmatrix}.$$

The hydrostatic stationary solutions of (4.5.16) satisfy

$$p_x = -\rho H_x. \tag{4.5.17}$$

For hydrostatic equilibrium, $u = 0$, and the profiles of pressure and density have to satisfy only condition (4.5.17), therefore there are infinite solutions that depend on an arbitrary function. Here we consider isothermal profiles, for which the gas temperature T is assumed to be constant.

For such perfect gas it is

$$p(x) = RT\rho(x)$$

then the differential relation (4.5.17) becomes

$$RT\frac{\rho_x}{\rho} = -H_x$$

with solution

$$\rho(x) = \bar{\rho}e^{-\frac{H(x)}{RT}}.$$

Choosing R and T such that $RT = 1$, a family of isothermal hydrostatic stationary solutions [21, 127] is given by

$$\rho^*(x) = C_1 e^{-H(x)} \geq 0; \quad p^*(x) = \rho^*(x) \geq 0; \quad u^*(x) = 0; \quad E^*(x) = \frac{p^*(x)}{\gamma - 1}. \tag{4.5.18}$$

In order to design numerical methods that preserve the stationary solutions of this family, the technique described in Remark 4.3.4 will be applied: given $U_i^n = [\rho_i^n, \rho_i^n u_i^n, E_i^n]^T$, the solution U_i^* of the family

$$\begin{aligned} \rho_i^*(x) &= \rho_i^n e^{-(H(x)-H(x_i))}, \\ p_i^*(x) &= \rho_i^n e^{-(H(x)-H(x_i))} \\ u_i^*(x) &= 0, \\ E_i^*(x) &= \frac{p_i^*(x)}{\gamma - 1}, \end{aligned} \tag{4.5.19}$$

i.e. the constant C_1 is chosen so that that $\rho^*(x_i) = \rho_i^n$.

Preservation of an isothermal stationary solution with linear H

Let us consider the Euler equations (4.5.16) in the space interval $[-1, 1]$ with gravitational potential $H(x) = x$ and initial condition

$$\rho(x, 0) = e^{-x}; \quad u(x, 0) = 0; \quad p(x, 0) = e^{-x}, \tag{4.5.20}$$

(see [21])

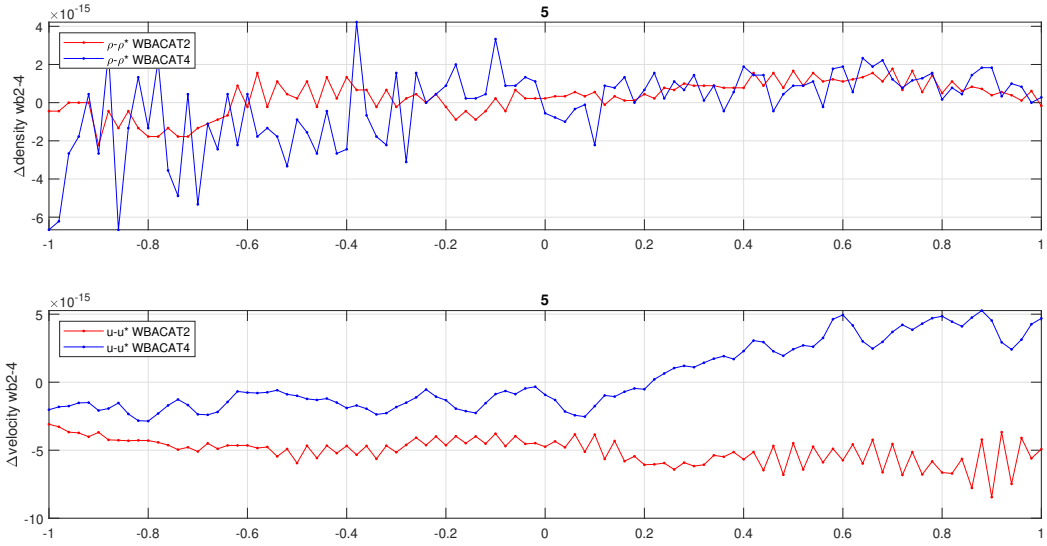

Figure 4.5.26: Test 4.5.4 (Preservation of an isothermal stationary solution with linear H). Differences between the exact and the numerical solutions obtained at time $t = 5$ with WBACAT2-4 for density (top) and velocity (bottom) using 100 mesh points and CFL= 0.8.

	ρ		u		p		E	
Points	Error	Order	Error	Order	Error	Order	Error	Order
50	2.16E-2	-	8.11E-2	-	9.74E-2	-	2.44E-2	-
100	5.35E-3	2.01	1.98E-2	2.03	2.42E-2	2.01	6.05E-3	2.01
200	1.33E-4	2.00	1.22E-3	2.02	1.51E-3	2.00	3.77E-4	2.00

Table 4.15: Test 4.5.4 (Preservation of an isothermal stationary solution with linear H). Errors in L^1-norm and convergence rate for ACAT2 at time $t = 5$.

The numerical solutions are computed on the interval $[-1, 1]$ using 100 mesh points and CFL= 0.8. As boundary conditions the exact stationary solution (4.5.20) is imposed at ghost points.

Figure 4.5.26 and Tables 4.15-4.17 show the differences between the numerical solutions at

	ρ		u		p		E	
Points	Error	Order	Error	Order	Error	Order	Error	Order
50	9.15E-6	-	3.45E-5	-	4.13E-5	-	1.03E-5	-
100	5.69E-7	4.01	2.11E-6	4.03	2.58E-6	4.00	6.44E-7	4.00
200	3.55E-8	4.00	1.30E-7	4.00	1.61E-7	4.02	4.02E-8	4.00

Table 4.16: Test 4.5.4 (Preservation of an isothermal stationary solution with linear H). Errors in L^1-norm and convergence rate for ACAT4 at time $t = 5$.

	WBACAT Order 2				WBACAT Order 4			
Points	ρ	u	p	E	ρ	u	p	E
50	1.79E-15	2.18E-15	2.11E-15	8.46E-15	2.05E-15	1.89E-15	1.43E-15	5.74E-15
100	1.54E-15	3.06E-15	1.27E-15	5.09E-15	1.13E-15	4.33E-15	2.49E-15	9.97E-15
200	3.53E-15	3.70E-15	9.89E-15	3.95E-15	8.19E-14	2.94E-15	1.17E-14	4.67E-14
400	5.14E-15	4.99E-15	1.45E-14	6.81E-15	6.89E-14	8.34E-14	1.97E-14	7.89E-14
800	1.57E-14	1.60E-14	4.25E-14	1.69E-14	6.05E-14	6.29E-14	4.72E-14	1.87E-13

Table 4.17: Test 4.5.4 (Preservation of an isothermal stationary solution with linear H). Errors in L^1-norm for WBACAT2-4 methods at time $t = 5$.

time $t = 5$ and stationary solution computed with, respectively, not well-balanced ACAT2-4 and well-balanced WBACAT2-4. As expected, the well-balanced methods capture the stationary solution with machine precision, confirmed by Tables 4.17. Nevertheless, the error in L^1-norm for the non well-balanced schemes decrease in accordance with the theoretical order, see in Tables 4.15-4.16.

As we have seen, the non well-balanced methods produce an error in accordance with their formal order of accuracy, while well-balanced methods capture the stationary solution with error of the order of machine precision. Now, we want to check what happens when a perturbation of the hydrostatic solution is used as initial condition and how the gravitational potential H plays an important role in the evolution of the perturbation.

Perturbation of an isothermal stationary solution with constant H

Let us consider the Euler equations (4.5.16) with a constant gravitational potential. This means that we are solving the Euler equations without source term. For the setting of this experiment we consider $H \equiv 1$, the interval $[-0.5, 1.5]$ and the initial conditions

$$\rho(x,0) = e^{-H} + 0.1e^{-200(x-0.5)^2}; \quad u(x,0) = 0; \quad p(x,0) = e^{-H} + 0.1e^{-200(x-0.5)^2} \tag{4.5.21}$$

which is a perturbation of the isothermal equilibrium (see Figure 4.5.27).

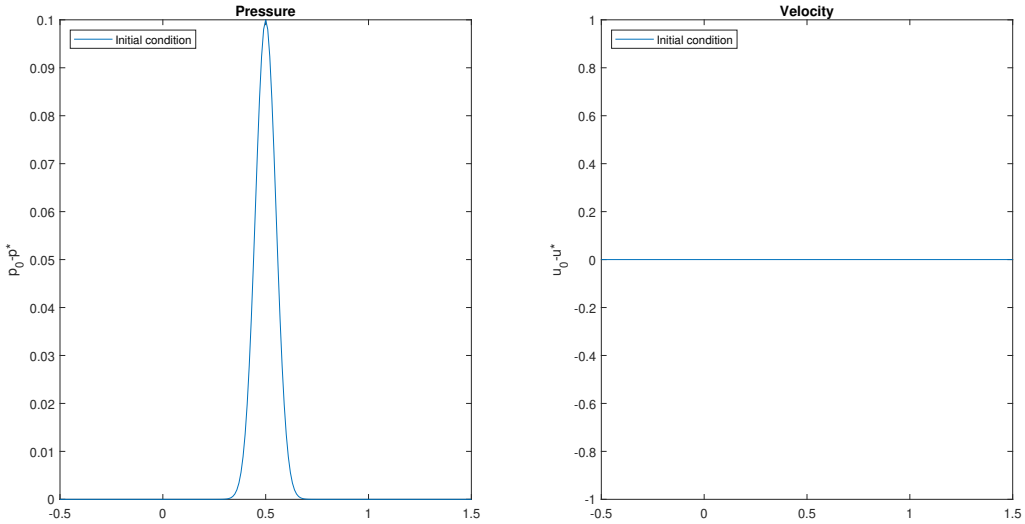

Figure 4.5.27: Test 4.5.4 (Perturbation of an isothermal stationary solution with constant H). Initial conditions: pressure (left); velocity (right).

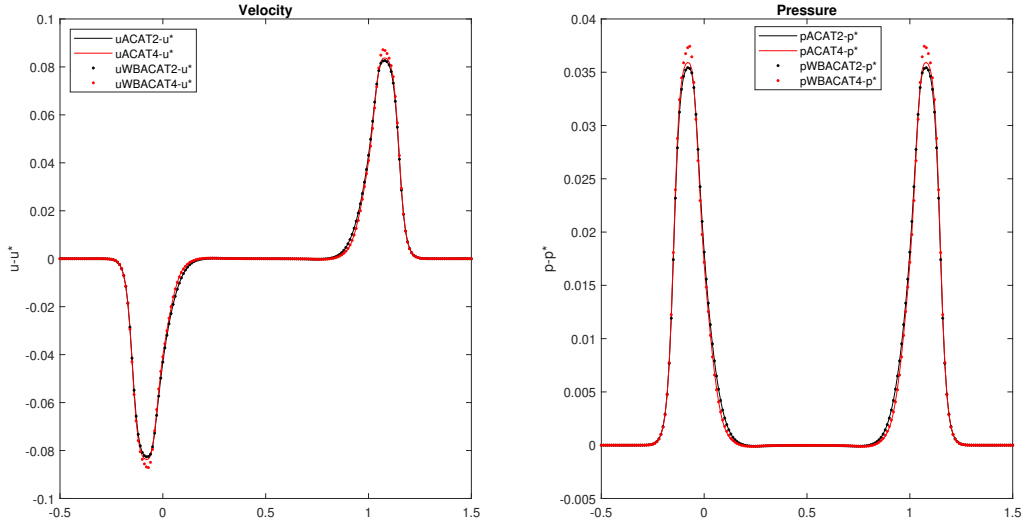

Figure 4.5.28: Test 4.5.4 (Perturbation of an isothermal stationary solution with constant H). Differences between the numerical solutions and the isothermal equilibrium obtained with ACAT2P and WBACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

The numerical solutions are computed on the interval $[-0.5, 1.5]$ using 200 mesh points and CFL$= 0.7$ at time $t = 0.5$. As boundary conditions the exact stationary solution is imposed at ghost points.

In Figure 4.5.28 the differences between the stationary solution and the numerical ones obtained with ACAT2P and WBACAT2P, $P = 1, 2$, is shown. Since the perturbation is not so small and the gravitational potential H is constant, all the methods, well-balanced

and non well-balanced, produce similar results in accordance with the theoretical order. Furthermore, due to the constant gravitational potential, we observe that the perturbation in position 0.5 is symmetrically split in two perturbation that evolve in both direction.

With the next experiments, following [4, 68, 117], we would like to check the perturbation evolution in presence of linear and non-linear gravitational potential. In particular, in the second case we will adopt a non-linear gravitational potential that presents a singularity near the computational domain.

Perturbation of an isothermal stationary solution with linear H

Let us consider the Euler equations (4.5.16) with a linear gravitational potential. For this experiment we consider $H(x) = x$, the interval $[-0.5, 1.5]$ and the initial conditions

$$\rho(x,0) = e^{-H(x)} + 0.1e^{-200(x-0.5)^2}; \quad u(x,0) = 0; \quad p(x,0) = e^{-H(x)} + 0.1e^{-200(x-0.5)^2} \tag{4.5.22}$$

which is a perturbation of the isothermal equilibrium as shown in Figure 4.5.29.

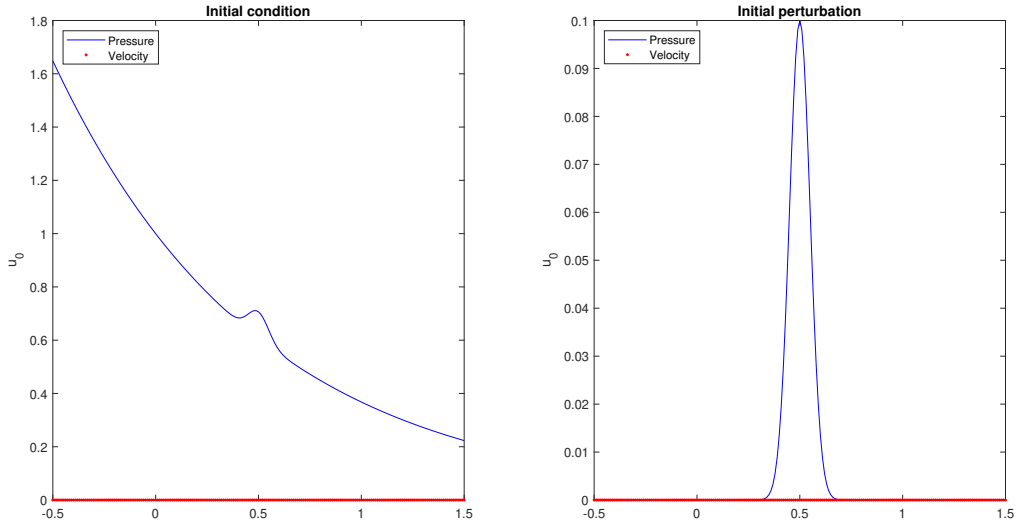

Figure 4.5.29: Test 4.5.4 (Perturbation of an isothermal stationary solution with linear H). Initial conditions (left); Initial perturbations (right).

The numerical solutions are computed in the interval $[-0.5, 1.5]$ using 200 mesh points and CFL= 0.7 at time $t = 0.5$. As boundary conditions the exact stationary solution is imposed at ghost points.

We remark that the overdetermined boundary conditions compatible with the stationary solutions can be imposed as far as the signal does not reach the boundary, which is the case

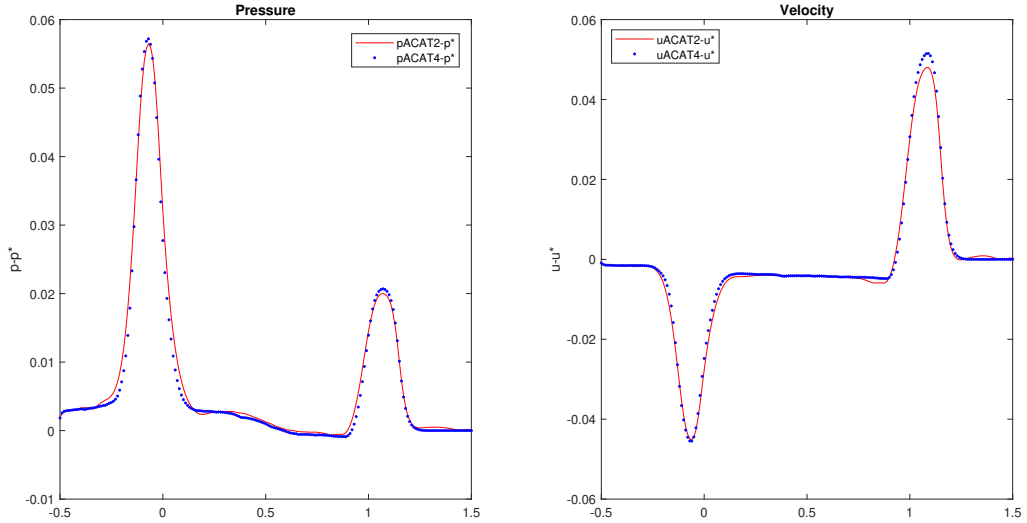

Figure 4.5.30: Test 4.5.4 (Perturbation of an isothermal stationary solution with linear H). Differences between the numerical solutions and the isothermal equilibrium obtained with ACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

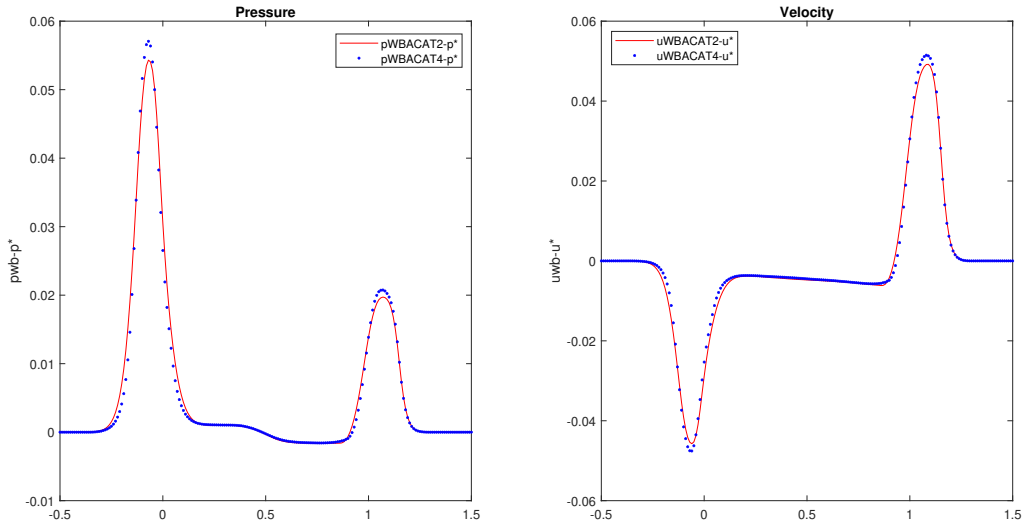

Figure 4.5.31: Test 4.5.4 (Perturbation of an isothermal stationary solution with linear H). Differences between the numerical solutions and the isothermal equilibrium obtained with WBACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

in the tests performed in this section.

Figures 4.5.30-4.5.31 show the difference between the stationary solution and the numerical ones obtained with ACAT2P and WBACAT2P, $P = 1, 2$. Since the perturbation is not small enough the difference between well-balanced and non well-balanced schemes is not so remarkable. Indeed, all the methods produce similar results in accordance with the theoretical order. Nevertheless, we could observe that, in the region in which the signal is not arrived, the well-balanced methods are able to preserve the stationary solution, behaviour

Points	Pressure							
	ACAT2		ACAT4		WBACAT2		WBACAT4	
125	7.78E-3	-	6.72E-3	-	4.10E-3	-	3.68E-3	-
250	3.75E-3	1.05	3.01E-3	1.16	1.88E-3	1.13	1.16E-3	1.66
500	1.55E-3	1.27	3.74E-4	3.01	7.16E-4	1.39	3.22E-4	1.85
1000	5.75E-4	1.43	2.48E-5	3.91	2.59E-4	1.47	2.11E-5	3.93

Table 4.18: Test 4.5.4 (Perturbation of an isothermal stationary solution with linear H). Errors in L^1-norm and convergence rates for pressure at time $t = 0.5$.

that is not followed by the non well-balanced schemes. Furthermore, due to the linear gravitational potential, we observe that the pressure perturbation in position 0.5 is split into two perturbations that evolve in both direction where the right signal amplitude decrease and the left one increases. Table 4.18 shows the error in L^1-norm and the numerical convergence rates. As expected, due to the adaptive order strategy and consequently to the smoothness indicators, a really fine mesh is required to obtain the theoretical order.

To complete this series of experiments concerning the perturbation of isothermal equilibrium, a non-linear gravitational potential H with a singularity close to the left boundary of the domain is adopted.

Perturbation of an isothermal stationary solution with non-linear H

Let us consider the Euler equations (4.5.16) with a non-linear gravitational potential. For this experiment we consider $H(x) = \frac{1}{x+0.7}$, the interval $[-0.5, 1.5]$ and the initial conditions

$$\rho(x,0) = e^{-H(x)} + 0.1e^{-200(x-0.5)^2}; \quad u(x,0) = 0; \quad p(x,0) = e^{-H(x)} + 0.1e^{-200(x-0.5)^2} \quad (4.5.23)$$

which is a perturbation of the isothermal equilibrium as exhibited in Figure 4.5.32.

The numerical solutions are computed on the interval $[-0.5, 1.5]$ using 200 mesh points and CFL= 0.7 at time $t = 0.5$. As boundary conditions the exact stationary solution is imposed at ghost points.

Figures 4.5.33-4.5.34 show the differences between the stationary solution and the numerical ones obtained with ACAT2P and WBACAT2P, $P = 1, 2$. Even if the perturbation is not too small, the difference between well-balanced and non well-balanced schemes is very remarkable. In fact, all the methods are able to evolve the perturbation, but we could observe that, in the region in which the signal has not arrived, the well-balanced methods are able

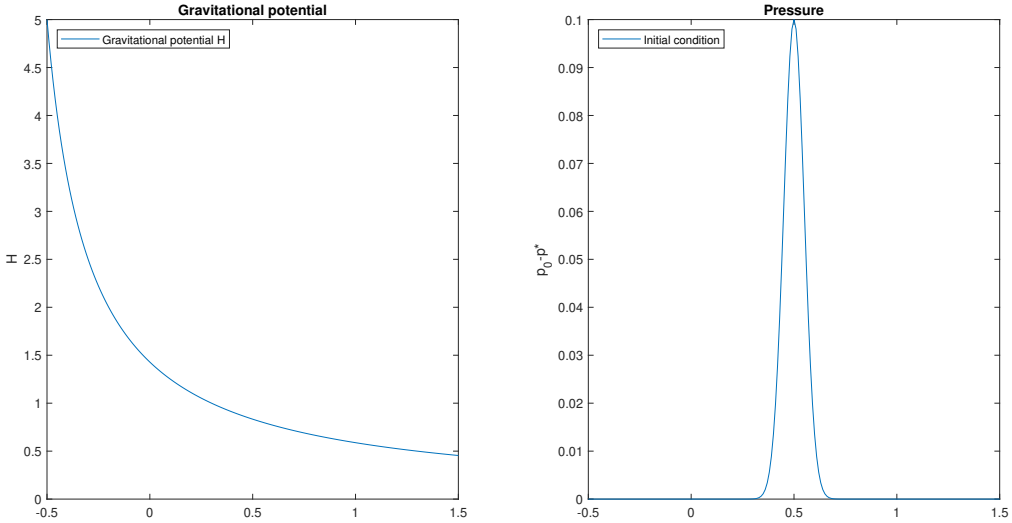

Figure 4.5.32: Test 4.5.4 (Perturbation of an isothermal stationary solution with non-linear H). Gravitational potential H (left); Initial condition (right).

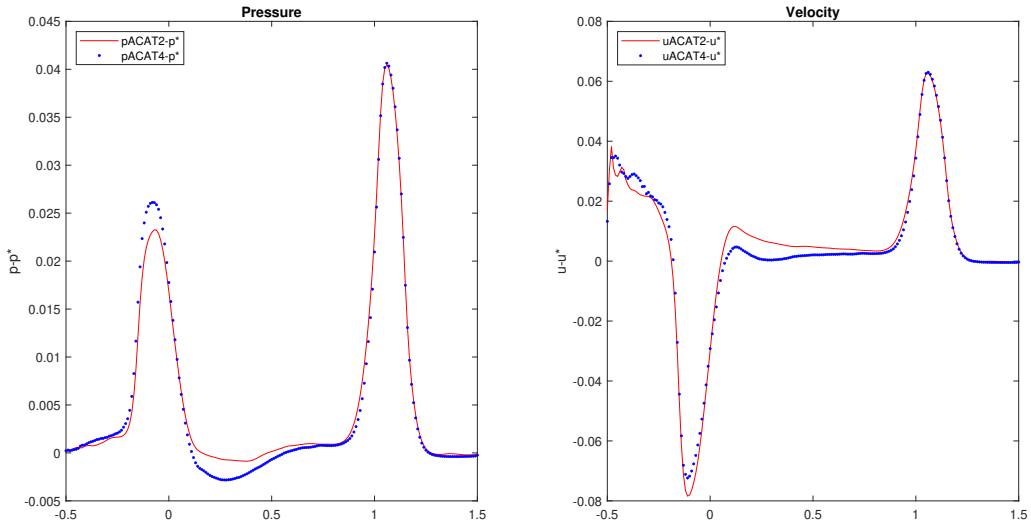

Figure 4.5.33: Test 4.5.4 (Perturbation of an isothermal stationary solution with non-linear H). Differences between the numerical solutions and the isothermal equilibrium obtained with ACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

to preserve the stationary solution, behaviour that is not followed by the non well-balanced schemes. Furthermore, due to the non-linear singular gravitational potential, we observe that the pressure perturbation in position 0.5 is split into two perturbations that evolve in both direction where the right signal amplitude are bigger than the left one. This behaviour is a direct consequence of the singularity in position $x = 0.7$.

We will finish the numerical experiments of this section with two very different tests. In the first one we consider a very small perturbation of the pressure where the signal amplitude is comparable with the acoustic regime; in the second one we consider the classic Shock Tube

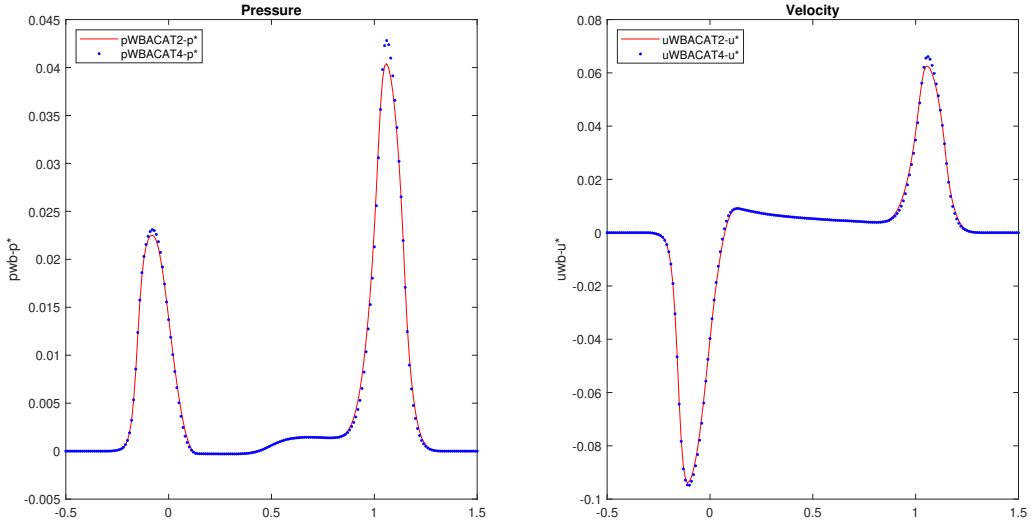

Figure 4.5.34: Test 4.5.4 (Perturbation of an isothermal stationary solution with non-linear H). Differences between the numerical solutions and the isothermal equilibrium obtained with WBACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

problem with constant and linear gravitational potential.

Acoustic regime

Let us consider the Euler equations (4.5.16) with a non-linear gravitational potential $H(x) = \frac{1}{x+0.7}$ on the interval $[-0.5, 1.5]$ and the initial conditions

$$\rho(x,0) = e^{-H(x)} + \varepsilon e^{-200(x-0.5)^2}; \quad u(x,0) = 0; \quad p(x,0) = e^{-H(x)} + \varepsilon e^{-200(x-0.5)^2} \tag{4.5.24}$$

which is a small perturbation of the isothermal equilibrium, in the acoustic regime with amplitude $\varepsilon = 10^{-6}$ (see Figure 4.5.35).

The numerical solutions are computed in the interval $[-0.5, 1.5]$ using 200 mesh points and CFL$= 0.7$ at time $t = 0.5$. As boundary conditions the exact stationary solution is imposed at ghost points.

Figures 4.5.36-4.5.37 show the difference between the stationary solution and the numerical ones obtained with ACAT2P and WBACAT2P, $P = 1, 2$, in the acoustic regime. In this case the non well-balanced reconstructions introduce errors that are much larger than the initial signal amplitude making it impossible to numerically reconstruct the signal. Furthermore, due to the non-linear singular gravitational potential, we observe that the well-balanced pressure perturbation in position 0.5 is split into two perturbations that evolve in both directions where the right signal amplitude are bigger than the left one. This

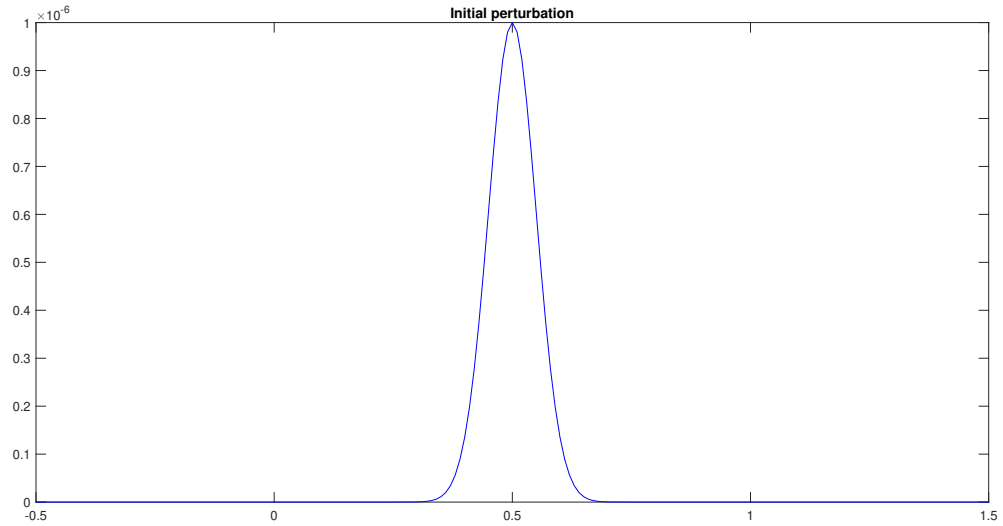

Figure 4.5.35: Test 4.5.4 (Acoustic regime). Pressure initial perturbation.

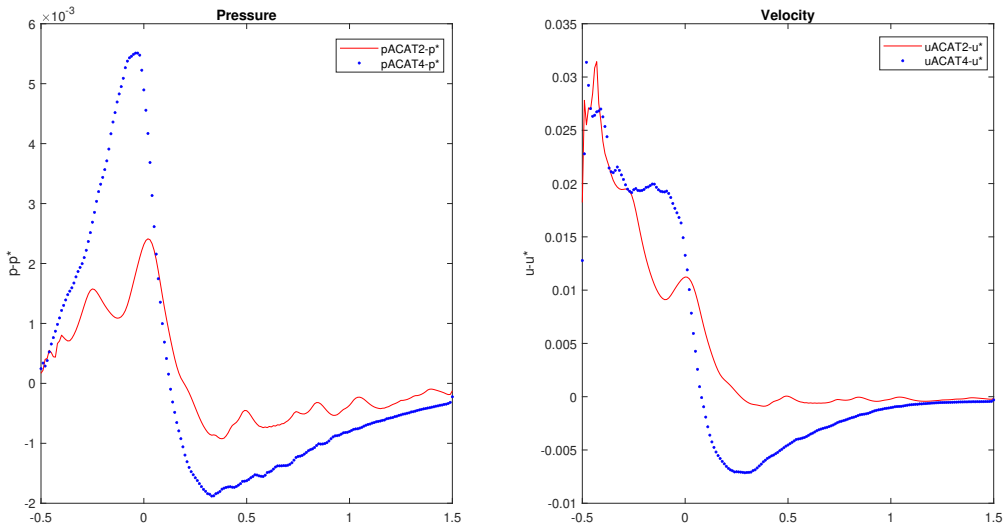

Figure 4.5.36: Test 4.5.4 (Acoustic regime). Differences between the numerical solutions and the isothermal equilibrium obtained with ACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL$= 0.7$ and a 200 mesh points.

behaviour is a direct consequence of the singularity in position $x = 0.7$.

Shock tube problem for Euler with gravity

In order to check if the well-balanced schemes give good results even when the solution to approximate is far from equilibrium, we consider the shock-tube problem see [24] in the space interval $[-0.5, 1.5]$ for (4.5.16) with the gravitational potential $H_1 \equiv 1$ and $H_2(x) = x$. The initial conditions are now

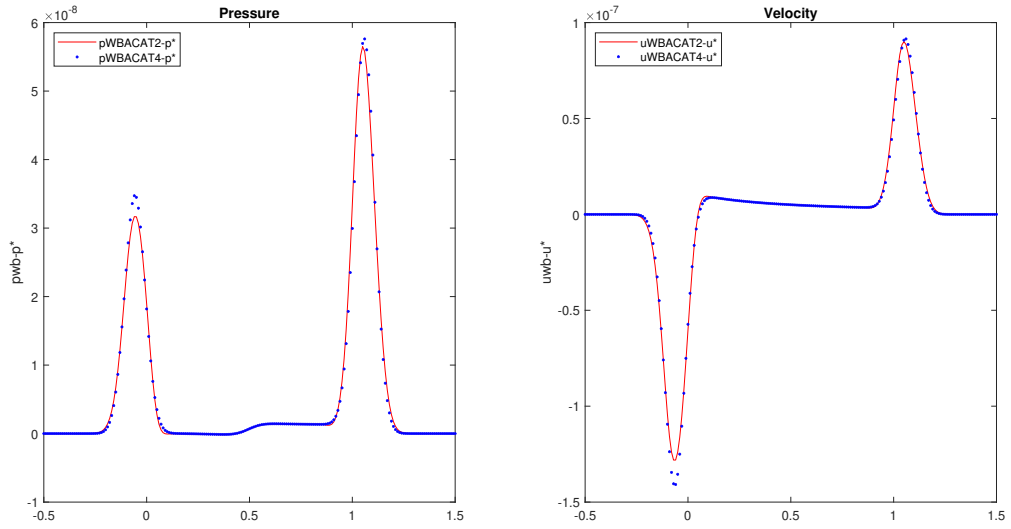

Figure 4.5.37: Test 4.5.4 (Acoustic regime). Differences between the numerical solutions and the isothermal equilibrium obtained with WBACAT2P, $P = 1, 2$, at time $t = 0.5$ with CFL= 0.7 and a 200 mesh points.

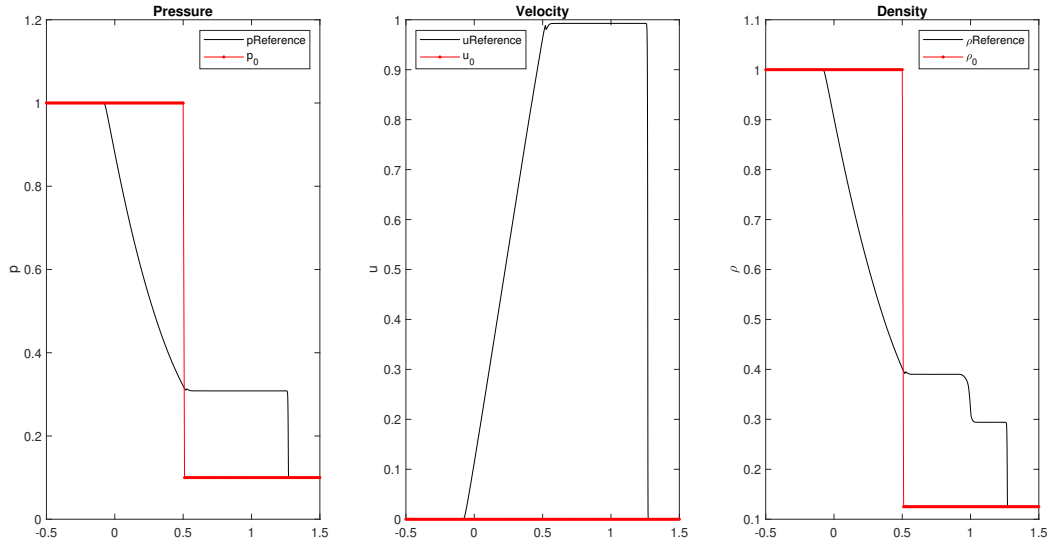

Figure 4.5.38: Test 4.5.4 (Shock tube problem for Euler with gravity). Initial conditions and reference solution computed with ACAT4 using 2000 mesh points and CFL= 0.5 for the gravitational potential $H_1 \equiv 1$.

$$[\rho(x,0),\ u(x,0),\ p(x,0)]^T = \begin{cases} [1,0,1]^T & \text{if} \quad x \leq \frac{1}{2} \\ [0.125,0,0.1]^T & \text{if} \quad x > \frac{1}{2}, \end{cases} \tag{4.5.25}$$

The numerical solutions are computed on the interval $[-0.5, 1.5]$ using 200 mesh points and CFL= 0.5. Dirichlet conditions are imposed at the boundaries. The reference solution is computed with ACAT4 using a 2000 mesh points.

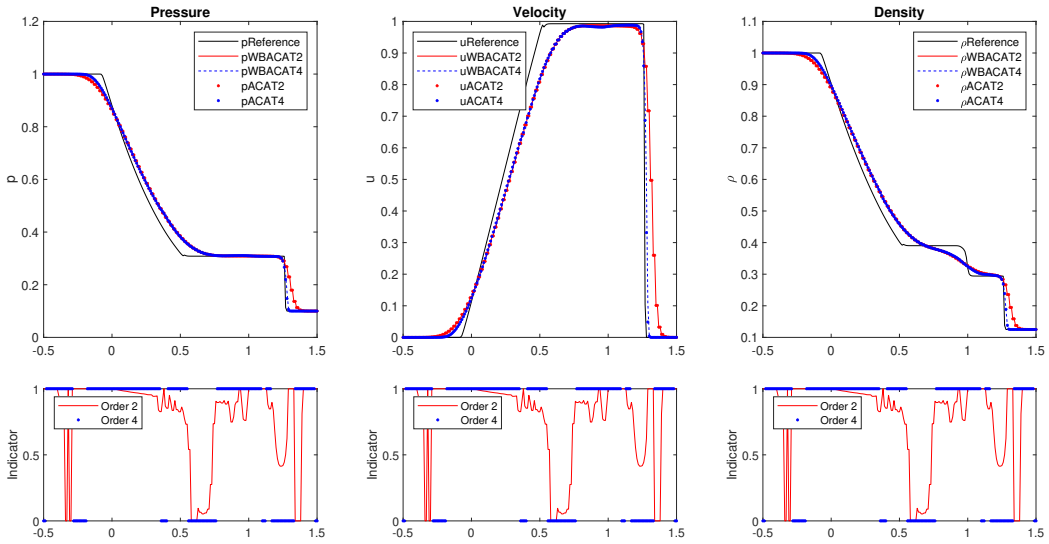

Figure 4.5.39: Test 4.5.4 (Shock tube problem for Euler with gravity). Reference and numerical solutions computed with well-balanced and non well-balanced ACAT2-4 at time $t = 0.5$ using 200 mesh points and CFL$= 0.5$: pressure (left), velocity (center) and density (right) with gravitational potential $H_1 \equiv 1$. The subframes show the indicators for ACAT2 and ACAT4. The reference solution is computed with ACAT4 using 2000 mesh points and CFL$= 0.5$

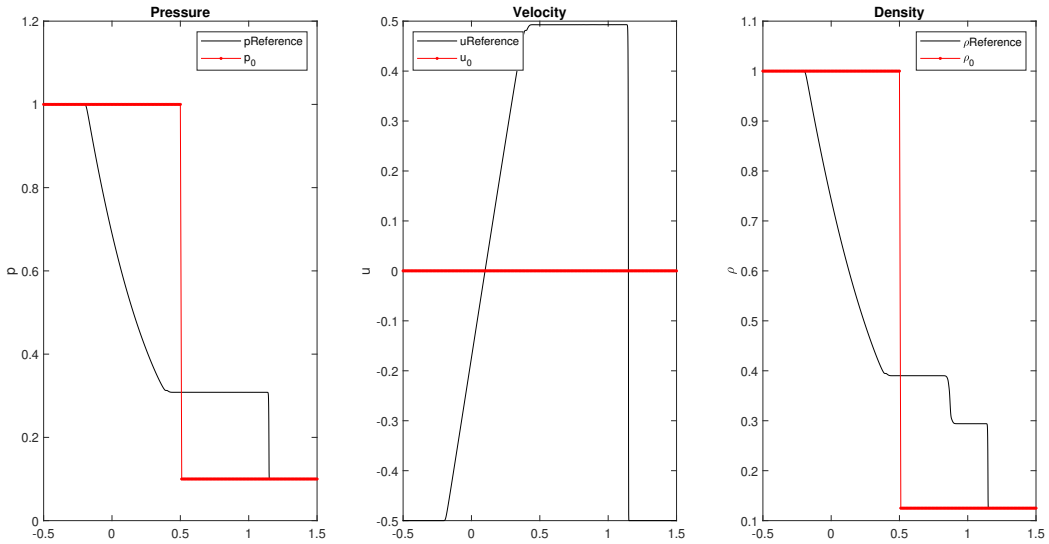

Figure 4.5.40: Test 4.5.4 (Shock tube problem for Euler with gravity). Initial conditions and reference solution computed with ACAT4 using 2000 mesh points and CFL$= 0.5$ for the gravitational potential $H_2(x) = x$.

Figures 4.5.39-4.5.41 show the reference and the numerical solutions at time $t = 0.5$ computed with ACAT2P and WBACAT2P, $P = 1, 2$, with gravitational potential H_1 and H_2. As it can be seen, the quality of the results obtained with well-balanced and not well-balanced methods are similar. Figures 4.5.38-4.5.40 exhibit the initial conditions and the

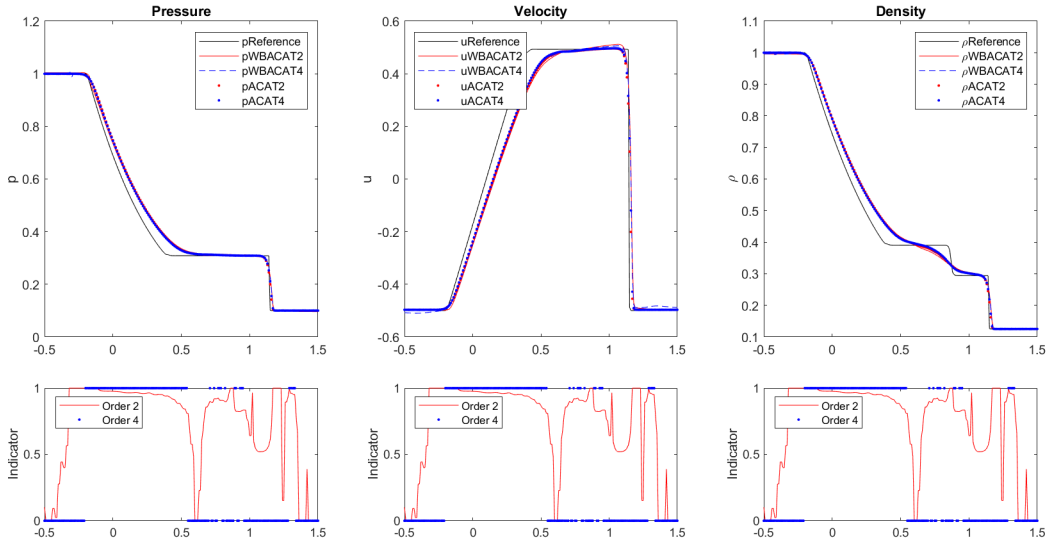

Figure 4.5.41: Test 4.5.4 (Shock tube problem for Euler with gravity). Reference and numerical solutions computed with well-balanced and non well-balanced ACAT2-4 at time $t = 0.5$ using 200 mesh points and CFL$= 0.5$: pressure (left), velocity (center) and density (right) with gravitational potential $H_2(x) = x$. The subframes show the indicators for ACAT2 and ACAT4. The reference solution is computed with ACAT4 using 2000 mesh points and CFL$= 0.5$

reference solution computed with ACAT4 adopting a 2000 mesh points. The subframes show the smoothness indicators for order 2 and 4 relative to ACAT2 and ACAT4.

From a careful evaluation it is noted that in Figure 4.5.39 the ACAT2 and WBACAT2 schemes have a behavior attributable to the Lax-Friedrichs method near the shocks. In our case, this phenomenon is mainly due to the second order fluxlimiter adopted. In fact, as can be seen from the subframes, the second order indicator is not able to capture the regularity in the right part of the shock, therefore applying the first order method which, in our case, is the Lax-Friedrichs scheme.

Chapter 5

2D Adaptive Compact Approximate Taylor Method for systems of balance law and well-balanced properties

The extension of ACAT scheme to non-linear two-dimensional systems of hyperbolic balance laws is shown in this section. For this reason, let us consider the 2D systems of hyperbolic balance laws so written:

$$U_t + f(U)_x + g(U)_y = S_1(U)H_x + S_2(U)H_y. \tag{5.0.1}$$

As we did for the 1D systems (4.0.1), let us introduce the functions $\mathcal{F}$ and $\mathcal{G}$ as:

$$\mathcal{F}(U)((x,y),t) = f(U(x,y,t)) - \int_{-\infty}^{x} S_1(U(\sigma))H_\sigma(\sigma,y)d\sigma; \tag{5.0.2}$$

$$\mathcal{G}(U)((x,y),t) = g(U(x,y,t)) - \int_{-\infty}^{y} S_2(U(\tau))H_\tau(x,\tau)d\tau, \tag{5.0.3}$$

assuming that the integrals are finite. Then, the identities

$$\mathcal{F}(U)_x = f(U(x,y,t))_x - S_1(U(x,y))H_x(x,y)$$

and

$$\mathcal{G}(U)_y = g(U(x,y,t))_y - S_2(U(x,y))H_y(x,y)$$

allow us one to write the 2D system of balance laws (5.0.1) in the equivalent conservative form

$$U_t + \mathcal{F}(U)_x + \mathcal{G}(U)_y = 0. \tag{5.0.4}$$

The idea is now extend the ACAT2P schemes to 2D systems written in the form (5.0.4).

5.1 2D Adaptive Compact Approximate Taylor Method for Systems of Balance Law

Following the notation used for the 2D systems of conservation laws, this multi-index notation will be used:

$$\mathbf{i} = (i_1, i_2) \in \mathbb{Z} \times \mathbb{Z},$$

and

$$\mathbf{0} = (0,0) \quad \mathbf{1} = (1,1) \quad \frac{\mathbf{1}}{\mathbf{2}} = (1/2, 1/2), \quad \mathbf{e}_1 = (1,0), \quad \mathbf{e}_2 = (0,1).$$

We consider Cartesian meshes with nodes

$$\mathbf{x}_\mathbf{i} = (i_1 \Delta x, i_2 \Delta y).$$

Using this notation, CAT2P methods for systems of balance laws can be extended as follows:

$$U_\mathbf{i}^{n+1} = U_\mathbf{i}^n + \frac{\Delta t}{\Delta x}\left[F^P_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + \widetilde{S}^P_{1,\mathbf{i}}\right] + \frac{\Delta t}{\Delta y}\left[G^P_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + \widetilde{S}^P_{2,\mathbf{i}}\right] \tag{5.1.1}$$

where the numerical fluxes $F^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ and the source terms $\widetilde{S}^P_{j,\mathbf{i}}$, $j = 1,2$ are computed using the values of the numerical solution $U_\mathbf{i}^n$ in the P^2-point stencil centered at $\mathbf{x}_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}} = ((i_1 + \frac{1}{2})\Delta x, (i_2 + \frac{1}{2})\Delta y)$

$$S^P_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}} = \{\mathbf{x}_{\mathbf{i}+\mathbf{j}}, \quad \mathbf{j} \in \mathfrak{I}_P\},$$

where

$$\mathfrak{I}_P = \{\mathbf{j} = (j_1, j_2) \in \mathbb{Z} \times \mathbb{Z}, \quad -P+1 \leq j_k \leq P, \quad k = 1,2\}.$$

See Figure 3.3.1 for an example with $P = 2$ in the one-dimensional case.

5.1.1 2D CAT2 for balance laws

Let us illustrate the extension of the 2D CAT procedure for systems of balance laws starting from the easiest case $P = 1$. In order to have a second order scheme, the quadrature formula used to compute integrals, dimension by dimension, in intervals of length $\Delta\sigma$ is the trapezoidal rule:

$$\int_{\sigma_i}^{\sigma_{i+1}} f(\sigma)\, d\sigma \approx \frac{\Delta\sigma}{2}\Big(f(\sigma_i) + f(\sigma_{i+1})\Big).$$

The numerical fluxes of second order are then as follows:

$$F^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \frac{1}{4}\left(f^{1,n+1}_{\mathbf{i},\mathbf{0}} + f^{1,n+1}_{\mathbf{i},\mathbf{e}_1} + f^n_{\mathbf{i}} + f^n_{\mathbf{i}+\mathbf{e}_1}\right), \tag{5.1.2}$$

$$G^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \frac{1}{4}\left(g^{1,n+1}_{\mathbf{i},\mathbf{0}} + g^{1,n+1}_{\mathbf{i},\mathbf{e}_2} + g^n_{\mathbf{i}} + g^n_{\mathbf{i}+\mathbf{e}_2}\right), \tag{5.1.3}$$

where

$$f^{1,n+1}_{\mathbf{i},\mathbf{j}} = f\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i},\mathbf{j}}\right),$$
$$g^{1,n+1}_{\mathbf{i},\mathbf{j}} = g\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i},\mathbf{j}}\right),$$

for $\mathbf{j} = \mathbf{0}, \mathbf{e}_1$ in the x-direction and $\mathbf{j} = \mathbf{0}, \mathbf{e}_2$ in the y-direction. Meanwhile, the source contribution is so computed:

$$\begin{aligned}
\widetilde{S}^1_{1,\mathbf{i}} = \frac{\Delta x}{8}\Big(&(S_1(U^n_{\mathbf{i}-\mathbf{e}_1}) + S_1(U^{1,n+1}_{\mathbf{i}-\mathbf{e}_1,\mathbf{0}}))H_x(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) + (S_1(U^n_{\mathbf{i}}) + S_1(U^{1,n+1}_{\mathbf{i}-\mathbf{e}_1,\mathbf{e}_1}))H_x(\mathbf{x}_{\mathbf{i}}) \\
&+(S_1(U^n_{\mathbf{i}}) + S_1(U^{1,n+1}_{\mathbf{i},\mathbf{0}}))H_x(\mathbf{x}_{\mathbf{i}}) + (S_1(U^n_{\mathbf{i}+\mathbf{e}_1}) + S_1(U^{1,n+1}_{\mathbf{i},\mathbf{e}_1}))H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1})\Big); \\
\widetilde{S}^1_{2,\mathbf{i}} = \frac{\Delta x}{8}\Big(&(S_2(U^n_{\mathbf{i}-\mathbf{e}_2}) + S_2(U^{1,n+1}_{\mathbf{i}-\mathbf{e}_2,\mathbf{0}}))H_y(\mathbf{x}_{\mathbf{i}-\mathbf{e}_2}) + (S_2(U^n_{\mathbf{i}}) + S_2(U^{1,n+1}_{\mathbf{i}-\mathbf{e}_2,\mathbf{e}_2}))H_y(\mathbf{x}_{\mathbf{i}}) \\
&+(S_2(U^n_{\mathbf{i}}) + S_2(U^{1,n+1}_{\mathbf{i},\mathbf{0}}))H_y(\mathbf{x}_{\mathbf{i}}) + (S_2(U^n_{\mathbf{i}+\mathbf{e}_2}) + S_2(U^{1,n+1}_{\mathbf{i},\mathbf{e}_2}))H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big).
\end{aligned} \tag{5.1.4}$$

On previous equations (5.1.4) we use the notation $U_{\mathbf{i}-\mathbf{j},\mathbf{j}}$ instead of $U_{i-m,(j-l,k)}$ to emphasize the different space positions according with dimension and notation introduced for CAT2P procedure for conservation laws. In particular, (l, k) are a single coordinate, respectively in dimension x or y but a double index is necessary to fix the reconstruction sides.

$U^{1,n+1}_{i,(j-l,k)}$ or $U^{1,n+1}_{(i-m,k),j-l}$ are the first order Taylor series computed as:

$$U^{1,n+1}_{\mathbf{i},\mathbf{j}} = U^{n}_{\mathbf{i},\mathbf{j}} + U^{(1)}_{\mathbf{i},\mathbf{j}}$$

for $\mathbf{j} = \mathbf{0}, \mathbf{e}_1$ in x dimension and $\mathbf{j} = \mathbf{0}, \mathbf{e}_2$ in y dimension where the first time derivatives $U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{j}}$ are so defined:

$$\begin{aligned}
U^{(1)}_{\mathbf{i},\mathbf{0}} = & - \frac{1}{\Delta x}\left(f(U^n_{\mathbf{i}+\mathbf{e}_1}) - f(U^n_{\mathbf{i}})\right) - \frac{1}{\Delta y}\left(g(U^n_{\mathbf{i}+\mathbf{e}_2}) - g(U^n_{\mathbf{i}})\right) \\
& + \frac{1}{2}\left(S_1(U^n_{\mathbf{i}})H_x(\mathbf{x}_{\mathbf{i}}) + S_1(U^n_{\mathbf{i}+\mathbf{e}_1})H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}) + S_2(U^n_{\mathbf{i}})H_y(\mathbf{x}_{\mathbf{i}})\right. \\
& + \left. S_2(U^n_{\mathbf{i}+\mathbf{e}_2})H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\right); \\
U^{(1)}_{\mathbf{i},\mathbf{e}_1} = & - \frac{1}{\Delta x}\left(f(U^n_{\mathbf{i}+\mathbf{e}_1}) - f(U^n_{\mathbf{i}})\right) - \frac{1}{\Delta y}\left(g(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - g(U^n_{\mathbf{i}+\mathbf{e}_1})\right) \\
& + \frac{1}{2}\left(S_1(U^n_{\mathbf{i}})H_x(\mathbf{x}_{\mathbf{i}}) + S_1(U^n_{\mathbf{i}+\mathbf{e}_1})H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}) + S_2(U^n_{\mathbf{i}+\mathbf{e}_1+})H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1})\right. \\
& + \left. S_2(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2})H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2})\right); \\
U^{(1)}_{\mathbf{i},\mathbf{e}_2} = & - \frac{1}{\Delta x}\left(f(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - f(U^n_{\mathbf{i}+\mathbf{e}_2})\right) - \frac{1}{\Delta y}\left(g(U^n_{\mathbf{i}+\mathbf{e}_2}) - g(U^n_{\mathbf{i}})\right) \\
& + \frac{1}{2}\left(S_1(U^n_{\mathbf{i}+\mathbf{e}_2})H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}) + S_1(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2})H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) + S_2(U^n_{\mathbf{i}})H_y(\mathbf{x}_{\mathbf{i}})\right. \\
& + \left. S_2(U^n_{\mathbf{i}+\mathbf{e}_2})H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\right).
\end{aligned}$$

The local expression of the second order numerical method is so get:

$$U^{n+1}_{\mathbf{i}} = U^n_{\mathbf{i}} + \frac{\Delta t}{\Delta x}\left[F^1_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + \widetilde{S}^1_{1,\mathbf{i}}\right] + \frac{\Delta t}{\Delta y}\left[G^1_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^1_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + \widetilde{S}^1_{2,\mathbf{i}}\right]. \tag{5.1.5}$$

5.1.2 2D CAT2P for balance laws

In order to be more readability let us introduce

$$\mathfrak{M}_P = \{\mathbf{j} \in \mathfrak{I} \quad \text{such that} \quad -\mathrm{P}+2 \leq \mathrm{j}_1 \leq \mathrm{P}\}$$

and

$$\mathfrak{N}_P = \{\mathbf{j} \in \mathfrak{I} \quad \text{such that} \quad -\mathrm{P}+2 \leq \mathrm{j}_2 \leq \mathrm{P}\}.$$

The following algorithm will be used to compute the numerical fluxes $F^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^P_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ and the source terms $\widetilde{S}^P_{j,\mathbf{i}}$, $j = 1, 2$ of 2D CAT2P method:

- Define

$$
\begin{aligned}
&F^{(0)}_{\mathbf{i},\mathbf{j}} := f(U^n_{\mathbf{i}+\mathbf{j}}), \quad \mathbf{j} \in \mathfrak{I}_P;\\
&G^{(0)}_{\mathbf{i},\mathbf{j}} := g(U^n_{\mathbf{i}+\mathbf{j}}), \quad \mathbf{j} \in \mathfrak{I}_P;\\
&I^{(0)}_{\mathbf{i},\mathbf{j}-\mathbf{e}_1,\mathbf{j}} := \Delta x \sum_{q=-P+1}^{P} a^{i_1,j_1}_{P,q} S_1(U^n_{\mathbf{i}+q\mathbf{e}_1}) H_x(x_{\mathbf{i}+q\mathbf{e}_1}), \quad \mathbf{j} \in \mathfrak{M}_P;\\
&I^{(0)}_{\mathbf{i},(-P+1,j_2),(-P+1,j_2)} = 0, \quad j_2 = -P+1,\ldots,P;\\
&I^{(0)}_{\mathbf{i},(-P+1,j_2),\mathbf{j}} = \sum_{s=-P+2}^{j_1} I^{(0)}_{\mathbf{i},(s-1,j_2),(s,j_2)} \quad \mathbf{j} \in \mathfrak{M}_P;\\
&J^{(0)}_{\mathbf{i},\mathbf{j}-\mathbf{e}_2,\mathbf{j}} := \Delta y \sum_{q=-P+1}^{P} a^{i_2,j_2}_{P,q} S_2(U^n_{\mathbf{i}+q\mathbf{e}_2}) H_y(x_{\mathbf{i}+q\mathbf{e}_2}), \quad \mathbf{j} \in \mathfrak{N}_P;\\
&J^{(0)}_{\mathbf{i},(j_1,-P+1),(j_1,-P+1)} = 0, \quad j_1 = -P+1,\ldots,P;\\
&J^{(0)}_{\mathbf{i},(j_1,-P+1),\mathbf{j}} = \sum_{s=-P+2}^{j_2} J^{(0)}_{\mathbf{i},(j_1,s-1),(j_1,s)} \quad \mathbf{j} \in \mathfrak{N}_P;
\end{aligned}
$$

- For $k = 1,\ldots,2P-1$:

  - Compute for all $\mathbf{j} \in \mathfrak{I}_P$

$$
\begin{aligned}
U^{(k)}_{\mathbf{i},\mathbf{j}} \;=\; & -A^{1,j_1}_P\Big(F^{(k-1)}_{\mathbf{i},(*,j_2)}, \Delta x\Big) + A^{1,j_1}_P\Big(I^{(k-1)}_{\mathbf{i},(-P+1,j_2),(*,j_2)}, \Delta x\Big)\\
& -A^{1,j_2}_P\Big(G^{(k-1)}_{\mathbf{i},(j_1,*)}, \Delta y\Big) + A^{1,j_2}_P\Big(J^{(k-1)}_{\mathbf{i},(j_1,-P+1),(j_1,*)}, \Delta y\Big).
\end{aligned}
$$

  - Compute for all $\mathbf{j} \in \mathfrak{I}_P$ and for all $r = -P+1,\ldots,P$

$$
U^{k,n+r}_{\mathbf{i},\mathbf{j}} = U^n_{\mathbf{i}+\mathbf{j}} + \sum_{m=1}^{k} \frac{(r\Delta t)^m}{m!} U^{(m)}_{\mathbf{i},\mathbf{j}}.
$$

  - Compute for all $\mathbf{j} \in \mathfrak{I}_P$ and for all $r = -P+1,\ldots,P$

$$
F^{k,n+r}_{\mathbf{i},\mathbf{j}} = f(U^{k,n+r}_{\mathbf{i},\mathbf{j}}) \quad \text{and} \quad G^{k,n+r}_{\mathbf{i},\mathbf{j}} = g(U^{k,n+r}_{\mathbf{i},\mathbf{j}}).
$$

– Compute for all $\mathbf{j} \in \mathfrak{M}_P$ and for all $r = -P+1, \ldots, P$

$$\begin{aligned} I^{k,n+r}_{\mathbf{i},\mathbf{j}-\mathbf{e}_1,\mathbf{j}} &= \Delta x \sum_{q=-P+1}^{P} a^{i_1,j_1}_{P,q} S_1(U^n_{\mathbf{i}+q\mathbf{e}_1}) H_x(x_{\mathbf{i}+q\mathbf{e}_1}); \\ I^{(k)}_{\mathbf{i},\mathbf{j}-\mathbf{e}_1,\mathbf{j}} &= A^{k,0}_P\Big(I^{k,*}_{\mathbf{i},\mathbf{j}-\mathbf{e}_1,\mathbf{j}}, \Delta t\Big). \end{aligned}$$

– Compute for all $\mathbf{j} \in \mathfrak{N}_P$ and for all $r = -P+1, \ldots, P$

$$\begin{aligned} J^{k,n+r}_{\mathbf{i},\mathbf{j}-\mathbf{e}_2,\mathbf{j}} &= \Delta y \sum_{q=-P+1}^{P} a^{i_2,j_2}_{P,q} S_2(U^n_{\mathbf{i}+q\mathbf{e}_2}) H_y(x_{\mathbf{i}+q\mathbf{e}_2}); \\ J^{(k)}_{\mathbf{i},\mathbf{j}-\mathbf{e}_2,\mathbf{j}} &= A^{k,0}_P\Big(J^{k,*}_{\mathbf{i},\mathbf{j}-\mathbf{e}_2,\mathbf{j}}, \Delta t\Big). \end{aligned}$$

– Compute

$$\begin{aligned} &F^{(k)}_{\mathbf{i},\mathbf{j}} = A^{k,0}_P\Big(F^{k,*}_{\mathbf{i},\mathbf{j}}, \Delta t\Big), \quad \mathbf{j} \in \mathfrak{I}_P; \\ &G^{(k)}_{\mathbf{i},\mathbf{j}} = A^{k,0}_P\Big(G^{k,*}_{\mathbf{i},\mathbf{j}}, \Delta t\Big), \quad \mathbf{j} \in \mathfrak{I}_P; \\ &I^{(k)}_{\mathbf{i},(-P+1,j_2),(-P+1,j_2)} = 0, \quad j_2 = -P+1, \ldots, P; \\ &I^{(k)}_{\mathbf{i},(-P+1,j_2),\mathbf{j}} = \sum_{s=-P+2}^{j_1} I^{(k)}_{\mathbf{i},(s-1,j_2),(s,j_2)} \quad \mathbf{j} \in \mathfrak{M}_P; \\ &J^{(k)}_{\mathbf{i},(j_1,-P+1),(j_1,-P+1)} = 0, \quad j_1 = -P+1, \ldots, P; \\ &J^{(k)}_{\mathbf{i},(j_1,-P+1),\mathbf{j}} = \sum_{s=-P+2}^{j_2} I^{(k)}_{\mathbf{i},(j_1,s-1),(j_1,s)} \quad \mathbf{j} \in \mathfrak{N}_P; \end{aligned}$$

- Compute

$$F^P_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}\mathbf{e}_1} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A^{0,\frac{1}{2}}_P\Big(F^{(k-1)}_{\mathbf{i},(*,0)}, \Delta x\Big); \tag{5.1.6}$$

$$G^P_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}\mathbf{e}_2} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A^{0,\frac{1}{2}}_P\Big(G^{(k-1)}_{\mathbf{i},(0,*)}, \Delta y\Big) \tag{5.1.7}$$

Once the algorithm has finished, the integrals will have already been computed and can be used to approximate the source terms as follows:

- For $k = 1, \ldots, 2P$ define

$$\mathcal{I}^{(k-1)}_{\mathbf{i},j_1} = \begin{cases} I^{(k-1)}_{\mathbf{i}-\mathbf{e}_1, j_1\mathbf{e}_1, (j_1+1)\mathbf{e}_1} & \text{if} \quad j_1 = -P+1, \ldots, 0; \\ I^{(k-1)}_{\mathbf{i}, (j_1-1)\mathbf{e}_1, j_1\mathbf{e}_1} & \text{if} \quad j_1 = 1, \ldots, P. \end{cases}$$

$$\mathcal{J}^{(k-1)}_{\mathbf{i},j_2} = \begin{cases} J^{(k-1)}_{\mathbf{i}-\mathbf{e}_2, j_2\mathbf{e}_2, (j_2+1)\mathbf{e}_2} & \text{if} \quad j_2 = -P+1, \ldots, 0; \\ J^{(k-1)}_{\mathbf{i}, (j_2-1)\mathbf{e}_2, j_2\mathbf{e}_2} & \text{if} \quad j_2 = 1, \ldots, P. \end{cases}$$

- Compute

$$\widetilde{S}^P_{1,\mathbf{i}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,\frac{1}{2}} \Big(\mathcal{I}^{(k-1)}_{\mathbf{i},*}, \Delta x \Big); \tag{5.1.8}$$

$$\widetilde{S}^P_{2,\mathbf{i}} = \sum_{k=1}^{2P} \frac{\Delta t^{k-1}}{k!} A_P^{0,\frac{1}{2}} \Big(\mathcal{J}^{(k-1)}_{\mathbf{i},*}, \Delta y \Big). \tag{5.1.9}$$

5.1.3 2D ACAT2P for balance laws

The extension of the adaptive CAT2P (5.0.1) is similar to conservation case (3.3.1). Following what has been done for ACAT2P in 1D we define $\mathcal{A}$, the indices set used to select the squared stencils according with the smoothness of numerical data, as:

$$\mathcal{A}_\mathbf{i} = \{p \in \{2, \ldots, P\} \text{ such that } \psi^p_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_1} \approx 1 \quad \text{and} \quad \psi^{\mathrm{p}}_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_2} \approx 1\} \tag{5.1.10}$$

where $\psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $\psi^p_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ are the smoothness indicators introduced in Subsection 3.2.2 computed direction by direction. Then the 2D adaptive CAT2P are so defined:

$$F^{A_\mathbf{i}}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \begin{cases} F^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{if } \mathcal{A}_\mathbf{i} = \emptyset; \\ F^{p_s}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{where } p_s = \max(\mathcal{A}_\mathbf{i}) \text{ otherwise;} \end{cases} \tag{5.1.11}$$

$$G^{A_\mathbf{i}}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \begin{cases} G^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{if } \mathcal{A}_\mathbf{i} = \emptyset; \\ G^{p_s}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{where } p_s = \max(\mathcal{A}_\mathbf{i}) \text{ otherwise;} \end{cases} \tag{5.1.12}$$

and

$$\widetilde{S}^{A_\mathbf{i}}_{j,\mathbf{i}} = \begin{cases} \widetilde{S}^*_{j,\mathbf{i}} & \text{if } \mathcal{A}_\mathbf{i} = \emptyset; \\ \widetilde{S}^{p_s}_{j,\mathbf{i}} & \text{where } p_s = \max(\mathcal{A}_\mathbf{i}) \text{ otherwise.} \end{cases} \tag{5.1.13}$$

$F^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^*_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$, and $\widetilde{S}^*_{j,\mathbf{i}}$ are the ACAT2 numerical fluxes and source terms given by 1D (4.2.5), (4.2.6); $F^{p_s}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^{p_s}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ and $\widetilde{S}^{p_s}_{j,\mathbf{i}}$ are the CAT2p_s numerical fluxes and source terms defined in (5.1.6), (5.1.8).

Using this notation, ACAT2P methods may be defined as follows:

$$U^{n+1}_{\mathbf{i}} = U^n_{\mathbf{i}} + \frac{\Delta t}{\Delta x}\left[F^{A_{\mathbf{i}}}_{\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^{A_{\mathbf{i}}}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + \widetilde{S}^{A_{\mathbf{i}}}_{1,\mathbf{i}}\right] + \frac{\Delta t}{\Delta y}\left[G^{A_{\mathbf{i}}}_{\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^{A_{\mathbf{i}}}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + \widetilde{S}^{A_{\mathbf{i}}}_{2,\mathbf{i}}\right] \tag{5.1.14}$$

Remark 5.1.1 *Observe that, since the smoothness indicators are computed dimension by dimension, a rectangular stencil*

$$S^{p_1,p_2}_{\mathbf{i}+\frac{\mathbf{1}}{\mathbf{2}}} = \{\mathbf{x}_{\mathbf{i},\mathbf{j}}, \quad i_1 - p_1 + 1 \le j_1 \le i_1 + p_1, \quad i_2 - p_2 + 1 \le j_2 \le i_2 + p_2\},$$

could be used to compute the numerical fluxes $F^{p_1}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^{p_2}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$. Unfortunately, a lot of modification are necessary to the quadrature rule increasing the computational cost without any immediately advantage.

5.2 2D Adaptive Well Balanced Compact Approximate Taylor Method for Balance Law

In the case of the well-balanced methods WBCAT2P, there is an important difference: if the algorithm described in Subsection 4.3 (adopting the 2D above notation) wants to be used, the first step, to update the numerical solution at the point $\mathbf{x}_\mathbf{i}$ at time t_n, would be find a solution of the problem

$$\begin{cases} f(U)_x + g(U)_y = S_1(U)H_x + S_2(U)H_y \\ U(\mathbf{x}_\mathbf{i}) = U^n_\mathbf{i}. \end{cases} \tag{5.2.1}$$

This problem is obviously much more difficult to solve, either exactly or numerically, than (4.3.6) since it is about a nonlinear PDE system instead an ODE system. Moreover in this case there may exist infinitely many stationary solutions satisfying the condition at only one point $\mathbf{x}_\mathbf{i}$: some extra conditions have to be imposed to determine one of them.

Nevertheless, if the stationary solutions to be preserved constitute a k-parameter family,

$$U^*(x, y; C_1, \ldots, C_k),$$

with $k < d$, then the numerical strategy described in Remark 4.3.4 can be followed: this strategy will be used in Subsection 5.3 to preserve a family of stationary solutions of the 2D Euler system with gravity.

Following the same idea of 1D well-balanced ACAT methods and 2D schemes for balance laws, let us introduce the functions $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ as:

$$\tilde{\mathcal{F}}(U)((x,y),t) = f(U(x,y,t)) - f(U^*(x,y)) - \int_{-\infty}^{x} (S_1(U(\sigma)) - S_1(U^*(\sigma)))H_\sigma d\sigma; \quad (5.2.2)$$

$$\tilde{\mathcal{G}}(U)((x,y),t) = g(U(x,y,t)) - g(U^*(x,y)) - \int_{-\infty}^{y} (S_2(U(\tau)) - S_2(U^*(\tau)))H_\tau d\tau, \quad (5.2.3)$$

assuming that the integrals are finite. Then, the identities

$$\mathcal{F}(U)_x = f(U(x,y,t))_x - f(U^*(x,y))_x - (S_1(U(x,y,t)) - S_1(U^*(x,y)))H_x(x,y)$$

and

$$\mathcal{G}(U)_y = g(U(x,y,t))_y - g(U^*(x,y))_y - (S_2(U(x,y,t)) - S_2(U^*(x,y)))H_y(x,y)$$

allow us to rewrite the 2D system of balance laws (3.3.1) in the equivalent conservative form

$$U_t + \tilde{\mathcal{F}}(U)_x + \tilde{\mathcal{G}}(U)_y = 0. \quad (5.2.4)$$

The main is extend the well-balanced strategy adopted for the one-dimensional case to 2D systems written in the form (5.2.4).

Remark 5.2.1 *Differently from the one-dimensional case, the two-dimensional case or more does not verify the, one-dimensional (direction by direction), stationary condition* $f(U)_x - S_1(U)H_x = 0$. *In fact, in general, it happens that* $f(U)_x - S_1(U)H_x = g(U)_y - S_2(U)H_y$ *that should be different from zero. Nevertheless,* $\tilde{\mathcal{F}}(U^*)_x + \tilde{\mathcal{G}}(U^*)_y = 0$

5.2.1 2D WBCAT2 for balance laws

The extension of the 2D well-balanced CAT procedure for systems of balance laws starts from the easiest case $P = 1$. In order to have a second order scheme, the quadrature formula used to compute integrals, dimension by dimension, in intervals of length $\Delta\sigma$ is the trapezoidal

rule:

$$\int_{\sigma_i}^{\sigma_{i+1}} f(\sigma)\,d\sigma \approx \frac{\Delta\sigma}{2}\Big(f(\sigma_i)+f(\sigma_{i+1})\Big).$$

The numerical well-balanced reconstructions of second order are then as follows:

$$F^1_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \frac{1}{4}\Big(f^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{0}} + f^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{e}_1} + f^n_{\mathbf{i}} + f^n_{\mathbf{i}+\mathbf{e}_1} - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})) - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}))\Big), \tag{5.2.5}$$

$$F^1_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_1} = \frac{1}{4}\Big(f^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{0}} + f^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{e}_1} + f^n_{\mathbf{i}-\mathbf{e}_1} + f^n_{\mathbf{i}} - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})) - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}))\Big), \tag{5.2.6}$$

$$G^1_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \frac{1}{4}\Big(g^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{0}} + g^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{e}_2} + g^n_{\mathbf{i}} + g^n_{\mathbf{i}+\mathbf{e}_2} - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})) - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big), \tag{5.2.7}$$

$$G^1_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_2} = \frac{1}{4}\Big(g^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{0}} + g^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{e}_2} + g^n_{\mathbf{i}} + g^n_{\mathbf{i}-\mathbf{e}_2} - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})) - 2f(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_2}))\Big), \tag{5.2.8}$$

where

$$f^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{j}} = f\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{j}}\right), \qquad f^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{j}} = f\left(U^n_{\mathbf{i}-\mathbf{e}_1+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{j}}\right),$$
$$g^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{j}} = g\left(U^n_{\mathbf{i}+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{j}}\right), \qquad g^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{j}} = g\left(U^n_{\mathbf{i}-\mathbf{e}_2+\mathbf{j}} + \Delta t U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{j}}\right),$$

for $\mathbf{j} = \mathbf{0}, \mathbf{e}_1$ in the x-direction and $\mathbf{j} = \mathbf{0}, \mathbf{e}_2$ in the y-direction. Meanwhile, the source contributions are computed direction by direction as follows:

$$\begin{aligned}\widetilde{S}^1_{1,\mathbf{i}} = \frac{\Delta x}{8}\ \Big[&\Big(S_1(U^n_{\mathbf{i}-\mathbf{e}_1}) + S_1(U^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{0}}) - 2S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1})\Big)H_x(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) \\ &+\Big(S_1(U^n_{\mathbf{i}}) + S_1(U^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{e}_1})2S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})\Big)H_x(\mathbf{x}_{\mathbf{i}}) \\ &+\Big(S_1(U^n_{\mathbf{i}}) + S_1(U^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{0}}) - 2S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})\Big)H_x(\mathbf{x}_{\mathbf{i}}) \\ &+\Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_1}) + S_1(U^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{e}_1}) - 2S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1})\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1})\Big];\end{aligned} \tag{5.2.9}$$

$$\begin{aligned}\widetilde{S}^1_{2,\mathbf{i}} = \frac{\Delta x}{8}\ \Big[&\Big(S_2(U^n_{\mathbf{i}-\mathbf{e}_2}) + S_2(U^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{0}}) - 2S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_2})\Big)H_y(\mathbf{x}_{\mathbf{i}-\mathbf{e}_2}) \\ &+\Big(S_2(U^n_{\mathbf{i}}) + S_2(U^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_2,\mathbf{e}_2})2S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})\Big)H_y(\mathbf{x}_{\mathbf{i}}) \\ &+\Big(S_2(U^n_{\mathbf{i}}) + S_2(U^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{0}}) - 2S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}})\Big)H_y(\mathbf{x}_{\mathbf{i}}) \\ &+\Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_2}) + S_2(U^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{e}_2}) - 2S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big];\end{aligned} \tag{5.2.10}$$

On previous equations (5.2.9)-(5.2.10) we use the notation $U_{\mathbf{i};\mathbf{i}-\mathbf{j},\mathbf{j}}$ for $U_{\mathbf{i};\mathbf{i}-m,(\mathbf{j}-l,k)}$ to emphasize the different space positions according with dimension and notation introduced for CAT2P procedure for conservation and balance laws. In particular, (l,k) are a single coordinate, respectively in dimension x or y but a double index is necessary to fix the reconstruction

sides.

$U^{1,n+1}_{\mathbf{i};\mathbf{i},(\mathbf{j}-l,k)}$ or $U^{1,n+1}_{\mathbf{i};(\mathbf{i}-m,k),\mathbf{j}-l}$ are the first order Taylor series computed as:

$$U^{1,n+1}_{\mathbf{i};\mathbf{i},\mathbf{j}} = U^{n}_{\mathbf{i},\mathbf{j}} + U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{j}}$$

$$U^{1,n+1}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{j}} = U^{n}_{\mathbf{i}-\mathbf{e}_1,\mathbf{j}} + U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{j}}$$

for $\mathbf{j} = \mathbf{0}, \mathbf{e}_1$ in x dimension and $\mathbf{j} = \mathbf{0}, \mathbf{e}_2$ in y dimension where the first time derivatives $U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{j}}$ and $U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{j}}$ are so defined:

$$
\begin{aligned}
U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{0}} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}+\mathbf{e}_1}) - f(U^n_{\mathbf{i}})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}+\mathbf{e}_2}) - g(U^n_{\mathbf{i}})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_x(\mathbf{x}_{\mathbf{i}}) + \Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_1}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}))\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}) \\
& + \Big(S_2(U^n_{\mathbf{i}}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_y(\mathbf{x}_{\mathbf{i}}) + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_2}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big]; \\
U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{e}_1} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}+\mathbf{e}_1}) - f(U^n_{\mathbf{i}})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - g(U^n_{\mathbf{i}+\mathbf{e}_1})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_x(\mathbf{x}_{\mathbf{i}}) + \Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_1}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}))\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}) \\
& + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_1}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1}) \\
& + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2})\Big]; \\
U^{(1)}_{\mathbf{i};\mathbf{i},\mathbf{e}_2} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - f(U^n_{\mathbf{i}+\mathbf{e}_2})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}+\mathbf{e}_2}) - g(U^n_{\mathbf{i}})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_2}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}) \\
& + \Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}))\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_1+\mathbf{e}_2}) \\
& + \Big(S_2(U^n_{\mathbf{i}}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_y(\mathbf{x}_{\mathbf{i}}) + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_2}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big]; \\
U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{0}} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}}) - f(U^n_{\mathbf{i}-\mathbf{e}_1})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}+\mathbf{e}_2-\mathbf{e}_1}) - g(U^n_{\mathbf{i}-\mathbf{e}_1})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}-\mathbf{e}_1}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}))\Big)H_x(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) \\
& + \Big(S_1(U^n_{\mathbf{i}}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_x(\mathbf{x}_{\mathbf{i}}) + \Big(S_2(U^n_{\mathbf{i}-\mathbf{e}_1}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}))\Big)H_y(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) \\
& + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_2-\mathbf{e}_1}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2-\mathbf{e}_1}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2-\mathbf{e}_1})\Big]; \\
U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{e}_1} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}}) - f(U^n_{\mathbf{i}-\mathbf{e}_1})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}+\mathbf{e}_2}) - g(U^n_{\mathbf{i}})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}-\mathbf{e}_1}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}))\Big)H_x(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) \\
& + \Big(S_1(U^n_{\mathbf{i}}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_x(\mathbf{x}_{\mathbf{i}}) \\
& + \Big(S_2(U^n_{\mathbf{i}+}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}}))\Big)H_y(\mathbf{x}_{\mathbf{i}}) + \Big(S_2(U^n_{\mathbf{i}+\mathbf{e}_2}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big)H_y(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2})\Big]; \\
U^{(1)}_{\mathbf{i};\mathbf{i}-\mathbf{e}_1,\mathbf{e}_2} = & - \frac{1}{\Delta x}\Big(f(U^n_{\mathbf{i}+\mathbf{e}_2}) - f(U^n_{\mathbf{i}+\mathbf{e}_2-\mathbf{e}_1})\Big) - \frac{1}{\Delta y}\Big(g(U^n_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}) - g(U^n_{\mathbf{i}-\mathbf{e}_1})\Big) \\
& + \frac{1}{2}\Big[\Big(S_1(U^n_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}))\Big)H_x(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}) \\
& + \Big(S_1(U^n_{\mathbf{i}+\mathbf{e}_2}) - S_1(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}))\Big)H_x(\mathbf{x}_{\mathbf{i}+\mathbf{e}_2}) + \Big(S_2(U^n_{\mathbf{i}-\mathbf{e}_1}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}))\Big)H_y(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1}) \\
& + \Big(S_2(U^n_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}) - S_2(U^*_{\mathbf{i}}(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2}))\Big)H_y(\mathbf{x}_{\mathbf{i}-\mathbf{e}_1+\mathbf{e}_2})\Big].
\end{aligned}
$$

The second order well-balanced numerical method is then defined as:

$$U_{\mathbf{i}}^{n+1} = U_{\mathbf{i}}^{n} + \frac{\Delta t}{\Delta x}\left[F^{1}_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^{1}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + \widetilde{S}^{1}_{1,\mathbf{i}}\right] + \frac{\Delta t}{\Delta y}\left[G^{1}_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^{1}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + \widetilde{S}^{1}_{2,\mathbf{i}}\right]. \quad (5.2.11)$$

5.2.2 2D Adaptive well-balanced CAT2P

The 2D adaptive well-balanced CAT2P scheme is similar to the non well-balanced method (5.1.14). In practise, following what has been done for 2D ACAT2P for conservation and balance laws, we define $\mathcal{A}$, the indices set used to select the squared stencils according with the smoothness of numerical data, as:

$$\mathcal{A}_{\mathbf{i}} \;=\; \{p \in \{2,\dots,P\} \text{ s.t. } \psi^{p}_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_1} \approx 1 \quad \text{and} \quad \psi^{\mathrm{p}}_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_2} \approx 1\} \quad (5.2.12)$$

where $\psi^{p}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $\psi^{p}_{\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ are the smoothness indicators introduced in Section 3.2.2 computed direction by direction. For this reason, the 2D adaptive well-balanced CAT2P writes as follows:

$$U_{\mathbf{i}}^{n+1} = U_{\mathbf{i}}^{n} + \frac{\Delta t}{\Delta x}\left[F^{A_{\mathbf{i}}}_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_1} - F^{A_{\mathbf{i}}}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} + \widetilde{S}^{A_{\mathbf{i}}}_{1,\mathbf{i}}\right] + \frac{\Delta t}{\Delta y}\left[G^{A_{\mathbf{i}}}_{\mathbf{i};\mathbf{i}-\frac{1}{2}\mathbf{e}_2} - G^{A_{\mathbf{i}}}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} + \widetilde{S}^{A_{\mathbf{i}}}_{2,\mathbf{i}}\right], \quad (5.2.13)$$

where

$$F^{A_{\mathbf{i}}}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} = \begin{cases} F^{*}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{if } \mathcal{A}_{\mathbf{i}} = \emptyset; \\ F^{p_s}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1} & \text{where } p_s = \max(\mathcal{A}_{\mathbf{i}}) \text{ otherwise;} \end{cases} \quad (5.2.14)$$

$$G^{A}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} = \begin{cases} G^{*}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{if } \mathcal{A}_{\mathbf{i}} = \emptyset; \\ G^{p_s}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2} & \text{where } p_s = \max(\mathcal{A}_{\mathbf{i}}) \text{ otherwise;} \end{cases} \quad (5.2.15)$$

and

$$\widetilde{S}^{A_{\mathbf{i}}}_{j\mathbf{i}} = \begin{cases} \widetilde{S}^{*}_{j,\mathbf{i}} & \text{if } \mathcal{A}_{\mathbf{i}} = \emptyset; \\ \widetilde{S}^{p_s}_{j,\mathbf{i}} & \text{where } p_s = \max(\mathcal{A}_{\mathbf{i}}) \text{ otherwise.} \end{cases} \quad (5.2.16)$$

Here, $F^{*}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^{*}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$, and $\widetilde{S}^{*}_{j,\mathbf{i}}$ are the ACAT2 numerical fluxes and source terms given by 1D (4.4.6),(4.4.7); $F^{p_s}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_1}$, $G^{p_s}_{\mathbf{i};\mathbf{i}+\frac{1}{2}\mathbf{e}_2}$ and $\widetilde{S}^{p_s}_{j,\mathbf{i}}$ are the ACAT2p_s numerical fluxes and source terms defined in 1D (4.3.18), (4.3.16).

5.3 Numerical experiments

Let us consider the 2D system of compressible Euler equations with a gravitational potential

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = -\rho H_x, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = -\rho H_y, \\ E_t + (u(E+p))_x + (v(E+p))_y = -\rho u H_x - \rho v H_y. \end{cases} \tag{5.3.1}$$

Here, ρ is the density; u, the velocity in $x-$direction; v, the velocity in $y-$direction; p, the pressure; E, the energy per unit volume excluding the gravitational energy; and $H(x, y)$, the gravitational potential [71]. The pressure is supposed to satisfy the equation of state

$$p = (\gamma - 1)\Big(E - \frac{1}{2}\rho(u^2 + v^2)\Big),$$

where γ is the ratio between specific heats at constant pressure and volume, which is taken to be constant. System (5.3.1) can be written in the form (5.0.1) with

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}, \quad f(U) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{bmatrix}, \quad g(U) = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{bmatrix}, \quad S(U) = \begin{bmatrix} 0 \\ -\rho H_x \\ -\rho H_y \\ -\rho u H_x - \rho v H_y \end{bmatrix},$$

$S(U) = S_1(U)H_x + S_2(U)H_y$ where

$$S_1(U) = \begin{bmatrix} 0 \\ -\rho \\ 0 \\ -\rho u \end{bmatrix}, \quad \text{and} \quad S_2(U) = \begin{bmatrix} 0 \\ 0 \\ -\rho \\ -\rho v \end{bmatrix}.$$

Hydrostatic stationary solutions satisfy

$$u = 0, \quad v = 0, \quad \nabla p = -\rho \nabla H.$$

So following the 1D cases [52, 69] two-parameter family of isothermal stationary solution is

given by

$$\rho^*(\mathbf{x}) = C_1 e^{-H(\mathbf{x})} \geq 0; \quad p^*(\mathbf{x}) = C_2 \rho^*(\mathbf{x_i}) \geq 0; \quad u^* = v^* = 0; \quad E^* = \frac{p^*}{\gamma - 1}. \tag{5.3.2}$$

Given

$$U_{\mathbf{i}}^n = [\rho_{\mathbf{i}}^n, \rho_{\mathbf{i}}^n u_{\mathbf{i}}^n, \rho_{\mathbf{i}}^n v_{\mathbf{i}}^n, E_{\mathbf{i}}^n]^T,$$

the stationary solution $U_{\mathbf{i}}^*$, selected applying the technique described in Remark 4.3.4, is given by

$$\rho_{\mathbf{i}}^*(\mathbf{x}) = \rho_{\mathbf{i}}^n e^{-(H(\mathbf{x})-H(\mathbf{x_i}))}; \; p_{\mathbf{i}}^*(\mathbf{x}) = \rho_{\mathbf{i}}^n e^{-(H(\mathbf{x})-H(\mathbf{x_i}))}; \; u_{\mathbf{i}}^* = v_{\mathbf{i}}^* = 0; \; E_{\mathbf{i}}^* = \frac{p_{\mathbf{i}}^*}{\gamma - 1}. \tag{5.3.3}$$

5.3.1 Preservation of a continuous stationary solution

Following [53, 69, 71] we consider Euler equations in the 2D domain $[0,1] \times [0,1]$ with two different gravitational potentials

$$H_1(x,y) = x + y, \quad H_2(x,y) = \frac{1}{\sqrt{(x-\frac{1}{3})^2 + (y+\frac{1}{2})^2}}$$

and initial conditions

$$\rho(\mathbf{x},0) = e^{-H(\mathbf{x})}; \quad p(\mathbf{x},0) = e^{-H(\mathbf{x})}; \quad u(\mathbf{x},0) = v(\mathbf{x},0) = 0. \tag{5.3.4}$$

	2D density					
	2D ACAT2		2D ACAT4		2D WBACAT2	2D WBACAT4
Points	Error	Order	Error	Order	Error	Error
20×20	4.87E-6	-	7.85E-9	-	2.72E-17	2.96E-18
40×40	1.91E-6	1.35	1.01E-9	2.95	2.19E-17	2.39E-18
80×80	5.62E-7	1.76	8.54E-11	3.56	1.82E-17	2.39E-18
160×160	1.43E-7	1.98	5.61E-12	3.93	2.37E-18	2.64E-18

Table 5.1: Test 5.3.1: (Preservation of a continuous stationary solution). 2D Euler equations with gravity and gravitational potential H_1. Errors in L^1-norm for density at time $t = 0.3$.

We solve numerically the equations using a $(21 \times 21)-$point mesh and CFL$= 0.9$. As boundary conditions the exact solution is imposed to all sides through the ghost points. Tables 5.1 and 5.2 exhibit the errors in L^1-norm for ACAT2P, WBACAT2P, $P = 1, 2$

	2D density					
	2D ACAT2		2D ACAT4		2D WBACAT2	2D WBACAT4
Points	Error	Order	Error	Order	Error	Error
20×20	3.85E-5	-	3.87E-5	-	2.50E-17	3.77E-17
40×40	1.58E-5	1.28	5.16E-6	2.91	3.23E-17	3.59E-17
80×80	4.78E-6	1.72	4.45E-7	3.53	3.33E-17	3.33E-17
160×160	1.23E-6	1.96	2.89E-8	3.95	3.15E-17	3.23E-17

Table 5.2: Test 5.3.1: (Preservation of a continuous stationary solution). 2D Euler equations with gravity and gravitational potential H_2. Errors in L^1-norm for density at time $t = 0.3$.

corresponding to $H = H_1$ and $H = H_2$ respectively. As it can be seen, the differences between the solutions given by well-balanced and no well-balanced methods are bigger for $H = H_2$: please note that, when linear potential H_1 is considered, the stationary solution is essentially 1D while this is not true for $H = H_2$. In this case, we observe that ACAT2-4 cannot produce accurate solutions as in previous case therefore requiring the use of well-balanced scheme to preserve the stationary isothermal equilibrium.

We have checked that all the methods achieve the expected order, nevertheless we would see what happens when a small perturbation of the stationary solution is applied. The idea is that only the well-balanced reconstruction should be able to maintain the hydrostatic profile and the non well-balanced one should introduce some spurious errors that may lead the smoothness indicators with a consequence order reduction in the adaptive strategy.

5.3.2 Perturbation of a continuous stationary solution

For this reason we consider now the Euler equations in the 2D domain $[0, 1] \times [0, 1]$ with the gravitational potential H_2 and an initial condition that represents a perturbation of the hydrostatic stationary considered in the previous test case:

$$\rho(\mathbf{x}, 0) = e^{-H(\mathbf{x})} + 0.008e^{-200(x-0.5)^2-200(y-0.5)^2}; \quad p(\mathbf{x}, 0) = e^{-H(\mathbf{x})} + 0.008e^{-200(x-0.5)^2-200(y-0.5)^2};$$
$$u(\mathbf{x}, 0) = v(\mathbf{x}, 0) = 0. \tag{5.3.5}$$

Table 5.3 shows errors in L^1-norm and convergence rates for the numerical solutions obtained with ACAT2P and WBACAT2P and the reference solution at time $t = 0.2$, with $P = 1, 2$. As happened for Shallow water, in case that a small perturbation of the stationary solution is considered as initial condition, the well-balanced schemes manage to capture the

	2D density							
	2D ACAT2		2D ACAT4		2D WBACAT2		2D WBACAT4	
Points	Error	Order	Error	Order	Error	Order	Error	Order
20×20	4.49E-5	-	5.93E-6	-	8.27E-6	-	4.28E-7	-
40×40	2.47E-5	0.86	1.71E-6	1.79	4.41E-6	0.91	4.92E-8	2.71
80×80	1.21E-5	1.03	4.37E-7	1.97	2.08E-6	1.08	7.05E-9	2.80
160×160	5.45E-6	1.15	9.85E-8	2.15	8.13E-7	1.36	9.47E-10	2.90
320×320	2.43E-6	1.17	2.13E-8	2.21	2.45E-7	1.73	1.22E-10	2.96

Table 5.3: Test 5.3.2: (Perturbation of a continuous stationary solution). 2D Euler equations with gravity and gravitational potential H_2. Errors in L^1-norm for density at time $t = 0.2$.

solution with a better accuracy than standard methods. This phenomena is shown on Tables 5.3.

5.3.3 Acoustic propagation

As last experiment we consider the Euler equations in the 2D domain $[0, 2] \times [0, 2]$ with the gravitational potential H_3,

$$H_3(x, y) = \frac{1}{\sqrt{(x - 0.4)^2 + (y + 0.1)^2}},$$

and an initial condition that represents a very small perturbation of the hydrostatic stationary considered in the previous test case:

$$\rho(\mathbf{x}, 0) = e^{-H(\mathbf{x})} + 0.000001e^{-200(x-1)^2-200(y-1)^2}; \quad p(\mathbf{x}, 0) = \rho(\mathbf{x}, 0); \quad u(\mathbf{x}, 0) = v(\mathbf{x}, 0) = 0. \tag{5.3.6}$$

Figures 5.3.2 shows the differences between the numerical solutions and the stationary solution computed at time $t = 0.75$ with WBACAT2 a using 101×101 mesh points and CFL$= 0.8$. As expected, the singularity of the gravitational potential modifies the thickness of the corona relative to the signal propagation, thinning it close to the singularity. H_3 has a singularity on position $(0.4, -0.1)$.

As we can see in Figure 5.3.3, the non well-balanced method ACAT2 is not able to capture the evolution of the wave generated by the initial perturbation, since the numerical errors are much bigger than the wave amplitude: one would need a space step at least two orders of magnitude lower in order to have a truncation error of the same order of the signal, making

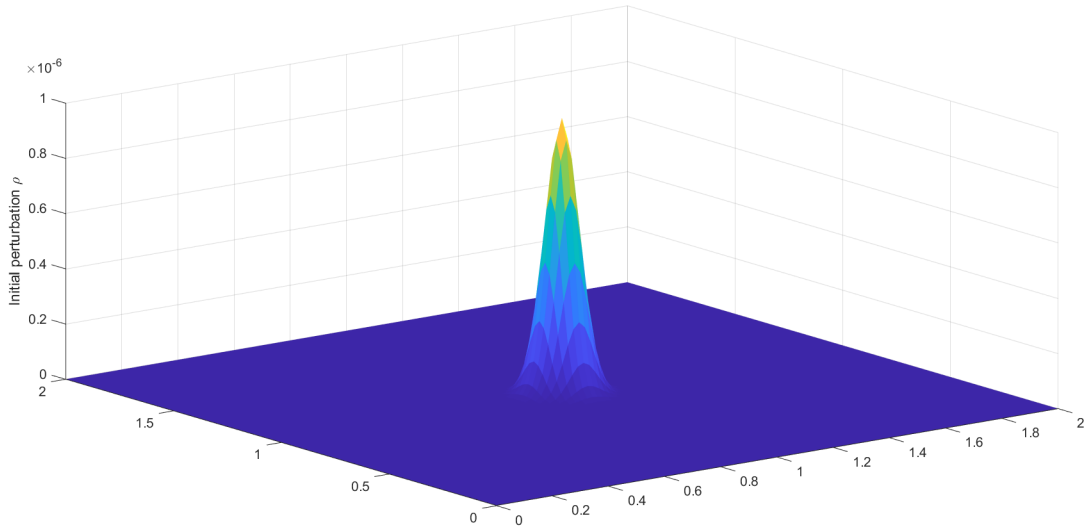

Figure 5.3.1: Test 5.3.3 (Acoustic propagation): 2D Euler equations with gravitational potential H_3. Initial perturbation using a 101×101 mesh points.

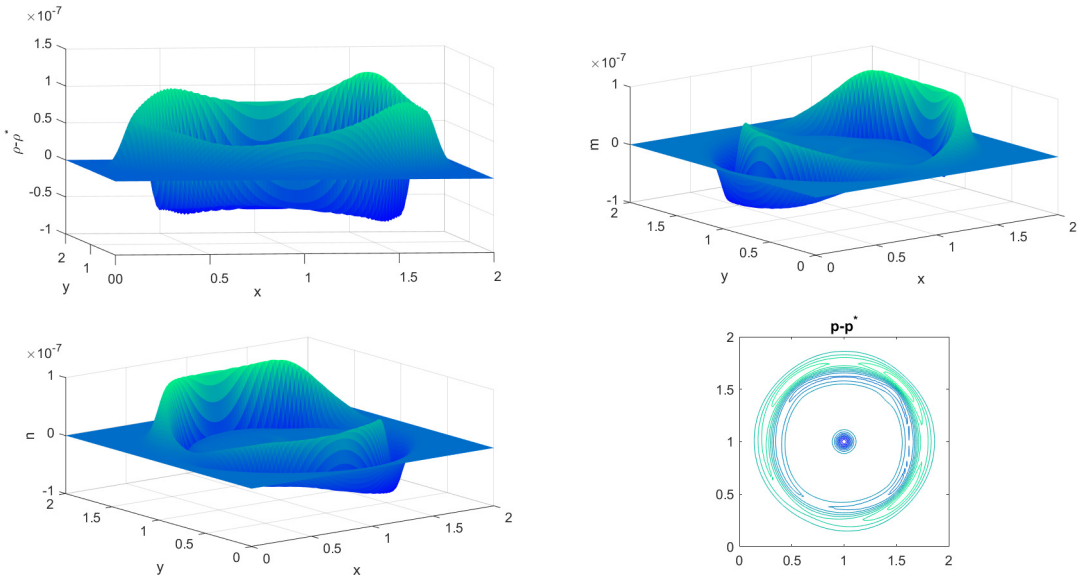

Figure 5.3.2: Test 5.3.3 (Acoustic propagation): 2D Euler equations with gravitational potential H_3. Differences between the numerical solutions and the stationary solution computed at time $t = 0.75$ with WBACAT2 using 101×101 mesh points and CFL$= 0.8$.

computation with non well-balanced method absolutely impractical.

Furthermore, in order to check the behaviour of the well-balanced reconstruction in presence of discontinuous initial condition we consider the Euler equations in the 2D domain $[0, 2] \times [0, 2]$ with the gravitational potential H_3,

$$H_3(x, y) = \frac{1}{\sqrt{(x - 0.4)^2 + (y + 0.1)^2}},$$

and an initial condition that represents a very small discontinuous perturbation of the hy-

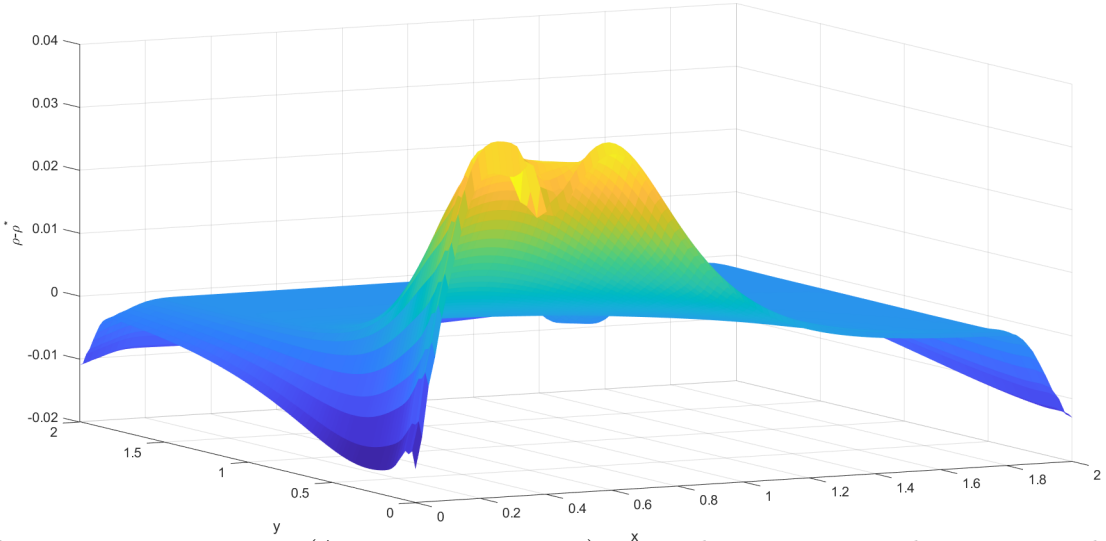

Figure 5.3.3: Test 5.3.3 (Acoustic propagation): 2D Euler equations with gravitational potential H_3. Difference between the density and the stationary solution computed at time $t = 0.75$ with ACAT2 using 101×101 mesh points and CFL$= 0.8$.

drostatic stationary considered in the previous test case:

$$\rho(\mathbf{x},0) = e^{-H(\mathbf{x})} + \begin{cases} 10^{-6} & \text{if } (x,y) \in [0.9,1.1]\times[0.9,1.1] \\ 0 & otherwise \end{cases} \tag{5.3.7}$$

$$p(\mathbf{x},0) = \rho(\mathbf{x},0); \quad u(\mathbf{x},0) = v(\mathbf{x},0) = 0. \tag{5.3.8}$$

Figures 5.3.4 shows the differences between the numerical solutions and the stationary solution computed at different times: initial condition (left-up); solution at time $t \approx 0.15$ (righ-up); solution at time $t \approx 0.35$ (left-down) and solution at time $t = 0.5$. with WBACAT2 a using 101×101 mesh points and CFL$= 0.7$. The gravitational potential H_3 has a singularity on position $(0.4, -0.1)$.

As we can see, the well-balanced method ACAT2 is able to capture the evolution of the wave generated by the initial perturbation even if a discontinuous pertirbation is considered.

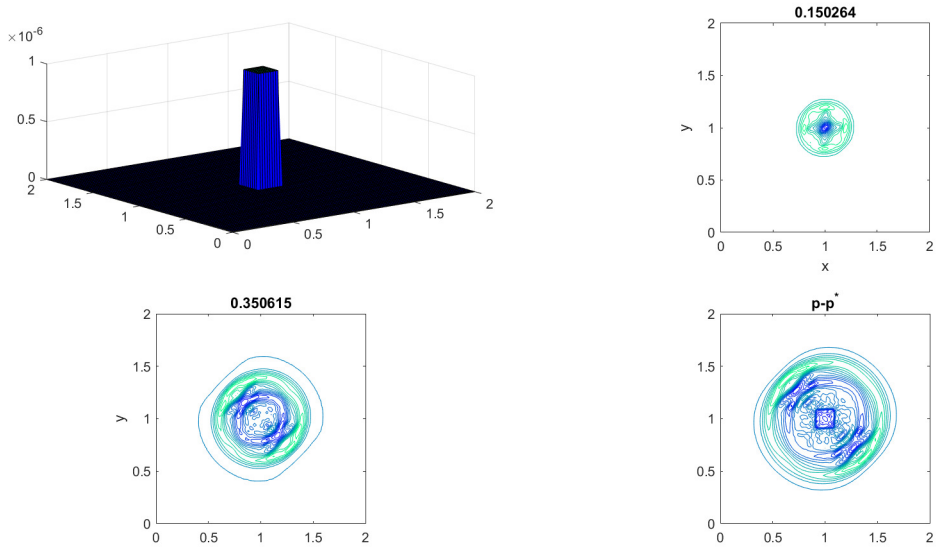

Figure 5.3.4: Test 5.3.3 (Acoustic propagation): 2D Euler equations with gravitational potential H_3. Numerical solutions for pressure obtained with WBACAT2 using a $101 \times 101-$ mesh points, CFL$= 0.7$ at different times: initial condition (left-up); solution at time $t \approx 0.15$ (righ-up); solution at time $t \approx 0.35$ (left-down) and solution at time $t = 0.5$.

Chapter 6

Semi-Implicit Exner model

The aim of this chapter is introduce an IMEX strategy to compute the sediment evolution [9, 18] in the Exner model of sediment transport in shallow water and improve both stability and efficiency. As expected, the velocity related to the sediment is very low respect to the free-surface wave. Unfortunately, as known, an explicit method implies a strong stability restriction due to the velocity of the free-surface wave. This restriction involves in a very long computation time that could be reduced neglecting the free-surface waves behaviour and looking at the sediment evolution. The objective is to drastically improve the efficiency in the computation of the evolution of the sediment by treating water waves implicitly, thus allowing much larger time steps than the one allowed by standard CFL condition on explicit schemes.

Recently, Garres-Díaz et al. (2022) proposed a semi-implicit $\Theta-$method approach for sediment transport models [44] by which, choosing $\theta > \frac{1}{2}$ in the semi-implicit method, an increasing in both efficiency and stability is obtained [22]. Differently from this paper we want check an IMEX strategy and a long time evolution such that the sediment initial dune moves 10 times the amplitude of the same. Furthermore, a reasonable approximation, under some conditions (see Section 6.3), consists in monitoring the sediment evolution on a sequence of quasi-stationary states increasing drastically the efficiency, in computation, of the method as it a scalar equation may be solved instead of a system of equations.

6.1 1D Exner Model

Let us consider the one-dimensional hyperbolic shallow water equation with bathymetry

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(\dfrac{q^2}{h} + \dfrac{g}{2}h^2\right)_x = -ghb_x, \end{cases} \tag{6.1.1}$$

where x makes reference to the axis of the channel and t is time; $q(x,t)$ and $h(x,t)$ represent

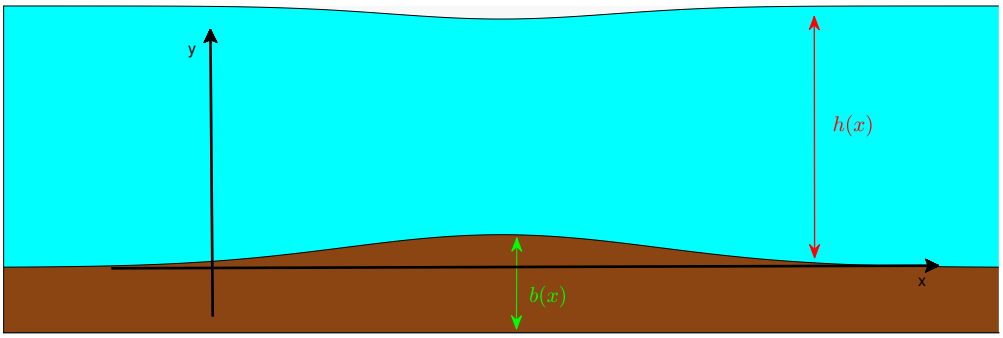

Figure 6.1.1: Shallow water equations: water-flow $h(x)$ and bottom topography $b(x)$.

the water-flow (discharge) and the thickness; g, the acceleration due to gravity; $b(x)$, the bottom topography; furthermore, the following relation holds $q(x,t) = h(x,t)u(x,t)$, with u the depth average horizontal velocity as shown in Figure 6.1.1.

The system of equations used in this work is obtained by coupling shallow water equation (6.1.1) and the sediment equation:

$$(z_b)_t + (q_b)_x = 0 \tag{6.1.2}$$

where $z_b(x,t)$ represents the height of sediment layer and $q_b(h,q)(x,t)$, the solid transport discharge, in our case computed by the Grass model [18, 50, 98]

$$q_b = \xi A_g u|u|^{m-1} \tag{6.1.3}$$

with $m \in [1,4] \cap \mathbb{N}$, $A_g \in]0,1[$ and $\xi = \dfrac{1}{1-\rho_0}$ where ρ_0 is the porosity of the sediment layer.

In this way, the Exner 1D system of balance laws is given by:

$$\begin{cases} h_t + q_x = 0, \\ q_t + (\frac{q^2}{h} + \frac{1}{2}gh^2)_x = -gh(b + z_b)_x, \\ (z_b)_t + (q_b)_x = 0. \end{cases} \tag{6.1.4}$$

Note that, if S is defined as $S(x,t) = b(x) + z_b(x,t)$, we have $\dfrac{\partial S}{\partial t} = \dfrac{\partial z_b}{\partial t}$, so system (6.1.4) could be rewritten as

$$\begin{cases} h_t + q_x = 0, \\ q_t + (\frac{q^2}{h} + \frac{1}{2}gh^2)_x = -gh(S)_x, \\ S_t + (q_b)_x = 0. \end{cases} \tag{6.1.5}$$

Observe that, system (6.1.4) can be written as a hyperbolic system with a non-conservative term

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = B(U)\frac{\partial U}{\partial x}, \tag{6.1.6}$$

where

$$U = \begin{bmatrix} h \\ q \\ S \end{bmatrix} \quad F = \begin{bmatrix} q \\ \frac{q^2}{h} + \frac{1}{2}gh^2 \\ q_b \end{bmatrix} \quad B(U) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -gh \\ 0 & 0 & 0 \end{bmatrix},$$

and q_b is given by eq. (6.1.3).

Given $J = \nabla_U F$ and $A(U) = J(U) - B(U)$, the system (6.1.6) could be rewritten as a non-conservative hyperbolic system

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = 0, \tag{6.1.7}$$

where,

$$A(U) = \begin{bmatrix} 0 & 1 & 0 \\ gh - u^2 & 2u & gh \\ \alpha & \beta & 0 \end{bmatrix},$$

in which $\alpha = \dfrac{\partial q_b}{\partial h} = -\xi A_g \dfrac{u^3}{h}$ and $\beta = \dfrac{\partial q_b}{\partial q} = 2\xi A_g \dfrac{u^2}{h}$ and $m = 3$ is selected in the grass equation (6.1.3). This system is strictly hyperbolic if and only if the characteristic

polynomial:

$$p_\lambda(\lambda) = -\lambda((u-\lambda)^2 - gh) + gh(\beta\lambda + \alpha)$$

has three distinct real roots $\lambda_1 < \lambda_2 < \lambda_3$.

Remark 6.1.1 *When* $A_g \to 0$,

$$\alpha = \frac{\partial q_b}{\partial h} = -\xi A_g \frac{u^3}{h} \to 0 \quad \text{and} \quad \beta = \frac{\partial q_b}{\partial q} = 2\xi A_g \frac{u^2}{h} \to 0.$$

Then, in the limit, $p_\lambda(\lambda) = -\lambda((u-\lambda)^2 - gh)$ *with distinct eigenvalues* $\lambda_\pm = u \pm \sqrt{gh}$ *and* $\lambda_0 = 0$, *where* $\lambda_- < \lambda_0 < \lambda_+$. *For this reason is not far suppose* $\lambda_2 \to 0$.

In our case, by assuming that the interaction between the water and the sediment is weak or $A_g << 1$, we are looking for the smallest eigenvalue (in absolute value). The wave speed of the sediment is much smaller than the water waves speed, therefore we assume that the eigenvalues corresponding to the sediment transport is the intermediate root λ_2 and that it is close to zero.

The idea behind this part is, under the hypothesis of $Fr = \frac{u}{\sqrt{gh}} << 1$, to use a semi-implicit method by which surface waves are treated implicitly while the sediment wave explicitly. The root λ_2, that could be found by a root finding algorithm such as Newton method etc., plays an important role since it could be used in a local Lax-Friedrichs flux based on the sediment wave, while the other waves are treated implicitly.

Finally, let us rewrite the 1D Exner model (6.1.4) in function of η where $\eta(x,t) = h(x,t) + b(x) + z_b(x,t)$ represents the elevation of the undisturbed water surface. In particular, system (6.1.5) with a non-conservative term will be:

$$\begin{cases} \eta_t + (q + q_b)_x = 0 \\ q_t + (qu)_x + gh(\eta)_x = 0 \\ (z_b)_t + (q_b)_x = 0 \end{cases} \tag{6.1.8}$$

(see Figure 6.1.2.)

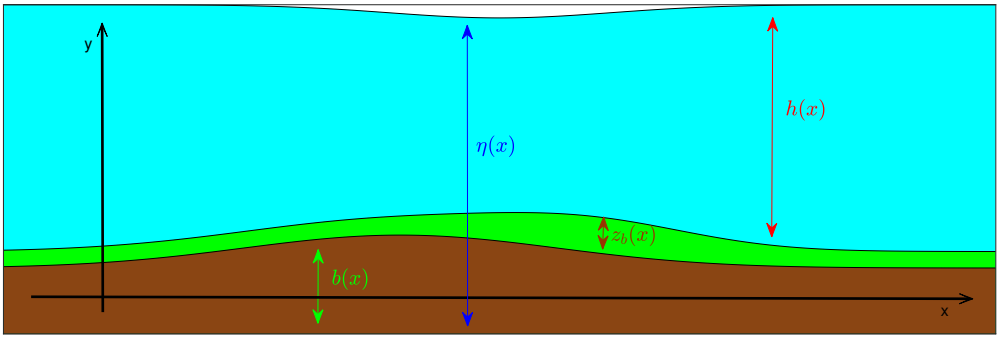

Figure 6.1.2: 1D Exner model: water surface $\eta(x)$; water-flow $h(x)$; sedimental layer $z_b(x)$ and bottom topography $b(x)$.

6.2 Semi implicit scheme

In this section, we will focus on the introduction of a scheme derived from an implicit treatment of the surface water waves, while the slow wave corresponding to the sediment evolution is treated explicitly. In particular, we will illustrate the first and second order semi-implicit schemes.

Let us consider a partition of the interval $[a, b]$ in cell defined by $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, with $i \in \mathbb{N}_0$. For the sake of simplicity, from now on we suppose that all the cells have the same length Δx and $x_i = a + i\Delta x$ are the cell centers. Let be Δt the time step such that $t^n = n\Delta t$.

Definitely, we denote by U_i^n an approximation on the mean value of U over cell I_i at time $t = t^n$,

$$U_i^n \cong \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U(x, t^n) dx.$$

6.2.1 First order scheme

Let us consider the system in non-conservative form (6.1.8). A semi-discrete in time first order semi-implicit scheme can be written as:

$$\begin{cases} \eta^{n+1} = \eta^n - \Delta t \hat{D}_x(q_b^n) - \Delta t D_x(q^{n+1}), \\ q^{n+1} = q^n - \Delta t \hat{D}_x(q^n u^n) - \Delta t g h^n D_x(\eta^{n+1}), \\ z_b^{n+1} = z_b^n - \Delta t \hat{D}_x(q_b^n), \end{cases} \tag{6.2.1}$$

where the differential operators D_x and $\hat{D}_x$ are respectively defined as:

- $D_x(f_i) = \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x}$, in which $f_{i\pm\frac{1}{2}}$ is suitably defined on cell edges;
- $\hat{D}_x(f_i) = \frac{F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}}{\Delta x}$, where $F_{i+\frac{1}{2}} = \frac{1}{2}\Big(f(v^-_{i+\frac{1}{2}}) + f(v^+_{i+\frac{1}{2}}) - \alpha_{i+\frac{1}{2}}\big(v^+_{i+\frac{1}{2}} - v^-_{i+\frac{1}{2}}\big)\Big)$ is the Rusanov flux and $\alpha_{i+\frac{1}{2}}$ is related to the eigenvalues of the explicit sub system. As we shall see, in our case $\alpha \approx |u| << \max(|\lambda_1|, |\lambda_3|)$.

For the sake of simplicity, let us rewrite systems (6.2.1) to distinguish explicit part from implicit one as:

$$\begin{cases} q^* = q^n - \Delta t \hat{D}_x(q^n u^n); \\ \eta^* = \eta^n - \Delta t \hat{D}_x(q^n_b) - \Delta t \hat{D}_x(q^*); \\ \eta^{n+1} = \eta^* + g\Delta t^2 D_x(h^n D_x(\eta^{n+1})); \\ q^{n+1} = q^* - \Delta t g h^n D_x(\eta^{n+1}); \\ z^{n+1}_b = z^n_b - \Delta t \hat{D}_x(q^n_b), \end{cases} \tag{6.2.2}$$

The procedure to solve system (6.2.1) and consequently (6.2.2) is:

1. solve explicitly $q^* = q^n - \Delta t \hat{D}_x(q^n u^n)$ as

$$q^*_i = q^n_i - \frac{\Delta t}{\Delta x}\Big(H^n_{i+\frac{1}{2}} - H^n_{i-\frac{1}{2}}\Big),$$

   where $H_{i\pm\frac{1}{2}}$ are the Rusanov flux, previously defined, related to the second equation;

2. solve explicitly $\eta^* = \eta^n - \Delta t \hat{D}_x(q^n_b) - \Delta t \hat{D}_x(q^*)$ as

$$\eta^*_i = \eta^n_i - \frac{\Delta t}{\Delta x}\Big(L^n_{i+\frac{1}{2}} - L^n_{i-\frac{1}{2}}\Big) - \frac{\Delta t}{\Delta x}\Big(Q^*_{i+\frac{1}{2}} - Q^*_{i-\frac{1}{2}}\Big),$$

   in which the spatial reconstruction $L_{i\pm\frac{1}{2}}$ and $Q^*_{i\pm\frac{1}{2}}$ are computed with the related Rusanov reconstruction;

3. fixed $k = g\Big(\frac{\Delta t}{\Delta x}\Big)^2$, solve implicitly $\eta^{n+1} = \eta^* + g\Delta t^2 D_x(h^n D_x(\eta^{n+1}))$ in the following way

$$\eta^{n+1}_i\Big(1 + k(h^n_{i+\frac{1}{2}} + h^n_{i-\frac{1}{2}})\Big) - \eta^{n+1}_{i+1} k h^n_{i+\frac{1}{2}} - \eta^{n+1}_{i-1} k h^n_{i-\frac{1}{2}} = \eta^*_i \quad \text{for all} \quad i = 1, \dots, N.$$

   This is an invertible tridiagonal linear system which can be solved to detect $\eta^{n+1} = [\eta^{n+1}_1, \dots, \eta^{n+1}_N]$;

4. solve explicitly $q^{n+1} = q^* - \Delta t g h^n D_x(\eta^{n+1})$

$$q_i^{n+1} = q_i^* - \frac{g\Delta t}{\Delta x} h_i^n \Big(\eta_{i+\frac{1}{2}}^{n+1} - \eta_{i-\frac{1}{2}}^{n+1}\Big),$$

where $\eta_{i\pm\frac{1}{2}}^{n+1} = \frac{1}{2}\Big(\eta_{i\pm1}^{n+1} + \eta_i^{n+1}\Big)$;

5. solve explicitly $z_b^{n+1} = z_b^n - \Delta t \hat{D}_x(q_b^n)$ as

$$z_{b_i}^{n+1} = z_{b_i}^n - \frac{\Delta t}{\Delta x}\Big(K_{i+\frac{1}{2}} - K_{i-\frac{1}{2}}\Big),$$

where $K_{i\pm\frac{1}{2}}$ are computed with the Rusanov flux, in general $K_{i\pm\frac{1}{2}} \neq L_{i\pm\frac{1}{2}}$;

6. compute $h_i^{n+1} = \eta_i^{n+1} - b_i - z_{b_i}^{n+1}$.

6.2.2 Second order scheme

Let us consider the system in non-conservative form (6.1.8), following the idea proposed in [8] in which an IMEX second order Runge-Kutta method is presented. Let us write system (6.1.8) in the partitioned ODE form in which the first component is related to the explicit terms and the second component to the implicit part:

$$U' = H(U, U). \tag{6.2.3}$$

In our case, $U = [\eta, q, z_b]^T$ and $H(U, U)$ is so define:

$$H(U,U) = \begin{bmatrix} -(q+q_b)_x \\ -(qu)_x - gh(\eta)_x \\ -(q_b)_x \end{bmatrix} \tag{6.2.4}$$

The partitioned ODE, distinguished between explicit and implicit part, will become:

$$H(U_E, U_I) = \begin{bmatrix} -\hat{D}_x((q_B)_E) & -D_x(q_I) \\ -\hat{D}_x((qu)_E) & -gh_E D_x(\eta_I) \\ -\hat{D}((q_b)_E) & \end{bmatrix}. \tag{6.2.5}$$

With this in mind, we consider 2 different Butcher tableau: one related to the explicit and one to the implicit reconstruction, as:

$$\begin{array}{c|cc} & 0 & \\ c & c & 0 \\ \hline & 1-\gamma & \gamma \end{array} \qquad \begin{array}{c|cc} \gamma & \gamma & \\ 1 & 1-\gamma & \gamma \\ \hline & 1-\gamma & \gamma \end{array} \tag{6.2.6}$$

where $\gamma = 1 - \frac{1}{\sqrt{2}}$ and $c = \frac{1}{2\gamma}$.

Remark 6.2.1 *Observe that the Butcher tableau* (6.2.6) *have identical b coefficients allowing only one final reconstruction.*

The general procedure to update the numerical solution from time t_n to t_{n+1} is as follows:

Step 1: Compute explicitly $U_E^{(i)}$ for all $i = 1, \ldots, s$ as:

$$U_E^{(i)} = U^n + \Delta t \sum_{j=1}^{i-1} a_{i,j}^E H(U_E^{(j)}, U_I^{(j)});$$

Step 2: Compute implicitly $U_I^{(i)}$ for all $i = 1, \ldots, s$ as:

$$U_I^{(i)} = U^n + \Delta t \sum_{j=1}^{i-1} a_{i,j}^I H(U_E^{(j)}, U_I^{(j)}) + a_{i,i}^I H(U_E^{(i)}, U_I^{(i)});$$

Step 3: $U^{n+1} = U_I(s)$.

In our case, applying the scheme defined by (6.2.6) we have:

1. $U_E^{(1)} = U^n$;
2. $U_I^{(1)} = U^n + \Delta t \gamma H(U_E^{(1)}, U_I^{(1)})$;
3. $U_E^{(2)} = U^n + \Delta t c H(U_E^{(1)}, U_I^{(1)})$;
4. $U_I^{(2)} = U^n + \Delta t (1-\gamma) H(U_E^{(1)}, U_I^{(1)}) + \Delta t \gamma H(U_E^{(2)}, U_I^{(2)})$;
5. $U^{n+1} = U_I^{(2)}$.

Remark 6.2.2 *Let observe that $U_E^{(2)}, U_I^{(2)}$ and $U_I^{(1)}$ have a common term, thus, step 3 and 4 may be rewritten as:*

$$U_E^{(2)} = (1 - \frac{c}{\gamma})U^n + \frac{c}{\gamma}U_I^{(1)};$$
$$U_I^{(2)} = (1 - \frac{1-\gamma}{\gamma})U^n + \frac{1-\gamma}{\gamma}U_I^{(1)} + \Delta t\gamma H(U_E^{(2)}, U_I^{(2)}).$$

6.3 Scalar Equation for 1D Exner Model

For weak coupling, i.e. for small values of the parameter A_g, the motion of the sediment takes place on a much longer time scale than surface waves. For such a reason, surface waves move over a bathimetry given by the bottom and the sediment, which is almost constant in time. We can therefore imagine that to detect the slow motion of the sediment, a reasonable approximation consists in monitoring the sediment motion on a sequence of quasi-stationary states. This is obtained by setting to zero the time derivative in the first two equations of the Exner model. Our starting point is therefore the following

$$\begin{cases} (q + q_b)_x = 0 \\ (qu)_x + gh(h + z_b + b)_x = 0 \\ (z_b)_t + (q_b)_x = 0 \end{cases} \tag{6.3.1}$$

From the first equation of (6.3.1) we get $q + q_b = Q$ hence, with the choice $m = 3$ and assuming $u > 0$,

$$h = \frac{Q}{u} - A_g u^2 \tag{6.3.2}$$

where, in this section for the sake of simplicity, we include the coefficient ξ in the parameter A_g.

Let us focus on the second equation of system (6.3.1) we have

$$(qu)_x + gh(h + z_b + b)_x = 0. \tag{6.3.3}$$

We can suppose $q \neq 0$. In fact, if $q = 0$ we directly obtain $u = 0$ and z_b constant in time. Then multiply (6.3.3) by $\frac{u}{q}$ we have:

$$\frac{u}{q}(qu)_x + g(h + z_b + b)_x = 0. \tag{6.3.4}$$

Let us define $G(u)$ a function such that

$$\frac{\partial G}{\partial x} = \frac{u}{q}(qu)_x; \tag{6.3.5}$$

$\frac{\partial G}{\partial x} = \frac{dG}{du}u_x$ hence $\frac{u}{q}(qu)_x = G'u_x$ and therefore

$$G'(u)u_x = \frac{u}{q}(q'u + q)u_x \quad \Rightarrow \quad G'(u) = \frac{q'}{q}u^2 + u. \tag{6.3.6}$$

As a consequence of the first equation of system (6.3.1) $q' = -q'_b = -3A_g u^2$, then G' takes the form

$$G'(u) = \frac{Q - 4A_g u^3}{Q - A_g u^3}u. \tag{6.3.7}$$

From equation (6.3.4) we obtain $G + g(h + z_b + b) = C$, where C is a constant, consequently $z_b = \frac{(C - G)}{g} - h - b$. Furthermore, from the third equation of (6.1.4) $z'_b u_t + q'_b u_x = 0$ which implies

$$z'_b = -\frac{G'}{g} + \frac{Q}{u^2} + 2Agu. \tag{6.3.8}$$

Finally, linking all the results obtained, we find the non-linear scalar equation

$$u_t + \lambda(u)u_x = 0 \tag{6.3.9}$$

where

$$\lambda(u) = \frac{3A_g u^2}{2A_g u + \frac{Q}{u^2} - \frac{G'}{g}}. \tag{6.3.10}$$

6.3.1 Second order numerical scheme

In order to solve numerically equation (6.3.9) we adopt the Lax-Wendroff scheme applied to equation in form (6.3.9) [96]. In particular, the Lax-Wendroff method is:

$$u(x_i, t_n + k) = u_i^n + k u_t \Big|_{x=x_i}^{t=t_n} + \frac{k^2}{2} u_{tt} \Big|_{x=x_i}^{t=t_n}. \tag{6.3.11}$$

u_t is immediately computed from the governing equation, $u_t = -\lambda(u)u_x$, while u_{tt} is defined as follow:

$$u_{tt} = \lambda(u)\Big[\lambda'(u)u_x^2 + \Big(\lambda(u)u_x\Big)_x\Big]. \tag{6.3.12}$$

The space derivative of u at position x_i at time t_n is computed with the second order central derivative:

$$u_x \Big|_{x=x_i}^{t=t_n} = \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x};$$

at the same time the space derivative of $\lambda(u)u_x$ at position x_i and time t_n is computed in the following way:

$$\Big(\lambda(u)u_x\Big)_x \Big|_{x=x_i}^{t=t_n} = \frac{\lambda_{i+\frac{1}{2}}^n \Big(u_{i+1}^n - u_i^n\Big) - \lambda_{i-\frac{1}{2}}^n \Big(u_i^n - u_{i-1}^n\Big)}{\Delta x^2} \tag{6.3.13}$$

with $\lambda_{i\pm\frac{1}{2}}^n = \frac{1}{2}\Big(\lambda(u_i^n) + \lambda(u_{i\pm1}^n)\Big)$.

At the end, the Lax-Wendroff scheme (6.3.11) becomes

$$\begin{aligned} u_i^{n+1} =& u_i^n - \Delta t \lambda(u_i^n)\Big(\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x}\Big) + \\ &+ \frac{\Delta t^2}{2}\lambda(u_i^n)\left[\lambda'(u_i^n)\Big(\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x}\Big) + \frac{\lambda_{i+\frac{1}{2}}^n \Big(u_{i+1}^n - u_i^n\Big) - \lambda_{i-\frac{1}{2}}^n \Big(u_i^n - u_{i-1}^n\Big)}{\Delta x^2}\right]. \end{aligned} \tag{6.3.14}$$

We plan to integrate the equation only in conditions in which the solution remains smooth, and for this reason we shall not use any limiter.

6.4 Numerical experiment

When large time steps are used, we should check whether we are able to correctly follow the sediment evolution even if the details on the fast water waves are lost. For this reason, we check the ability of the scheme, presented in previous section, to compute the bathymetry evolution. The main purpose is increase the CFL-condition as much as possible in order to reduce the computational cost using the IMEX strategy described before.

6.4.1 1D Exner test

In this section we will compare the solutions obtained by the second order explicit scheme, first and second order semi-implicit schemes applied to system (6.1.8) and first and second order explicit scheme applied to the scalar equation (6.3.9). With this purpose in mind, for the scalar equation (6.3.9), explicit scalar schemes require a standard CFL condition as $\text{CFL}_{scal} = 0.9$; for the second order explicit scheme applied to (6.1.8) we use $\text{CFL}_{expl} = 0.4$; finally, for the semi-implicit methods a larger CFL condition could be used, however, since the term qu (6.2.1) is treated explicitly, the semi-implicit CFL condition could not be arbitrary larger and a material CFL condition must be satysfied. In our case $\text{CFL}_{IMEX} = 15$ is adopted. The common settings of this experiment are: $[-2, 4]$ the interval; $A_g = 0.1$ as we have said only in section 6.3 ξ is included in A_g; $\rho_0 = 0.2$; $t_{fin} = 1400$; and, since in Section 6.3 all the variables are written depending on the velocity u, initial conditions are so set: $b(x) = 0$, $h_0(-2) = 0.5$,

$$u_0(x) = 0.1 + 0.006e^{-\frac{(x-0.4)^2}{0.4^2}} \tag{6.4.1}$$

and $z_b(-2) = 0.1$. The constant Q is obtained through $Q = q_0(-2) + q_b(-2)$; while C is computed as $C = G(u_0(-2)) + g(h_0(-2) + z_b(-2) + b(-2))$ where G is a solution of (6.3.7). Free boundary conditions for left and right part are imposed at ghost points.

Figure 6.4.1 shows the initial condition of the height sediment layer z_b (center), the thickness η (top) and velocity u (down). The initial condition of thickness marks out by equation (6.3.2) while the initial condition of sediment layer comes out from $z_b = \frac{C-G(u_0)}{g} - h_0(x) - b(x)$. Figure 6.4.2 exhibits the time evolution of thickness η (top); sediment layer z_b (center); and velocity u (down) at final time $t = 1400$. The zoom of critical parts are shown in Figure 6.4.3. Table 6.1 proves that all the methods are able to keep the expected order refining the mesh-grid. The final time is set in order to not introduce a shock on velocity,

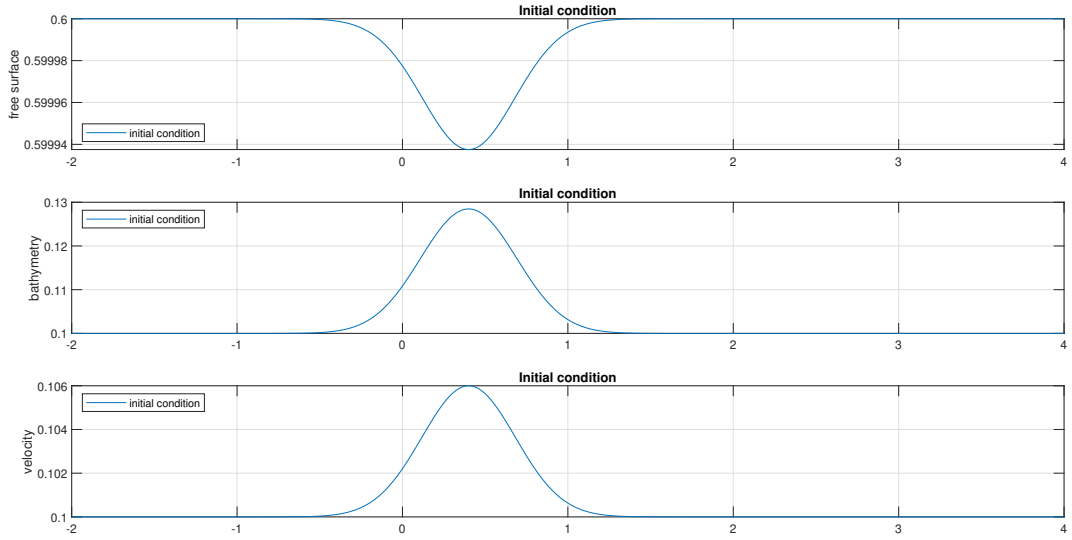

Figure 6.4.1: Test 6.4.1: (1D Exner test). Initial condition of thickness (top), sediment layer (center) and velocity (down) for the Exner model on the interval $[-2, 4]$ using a 200−mesh points.

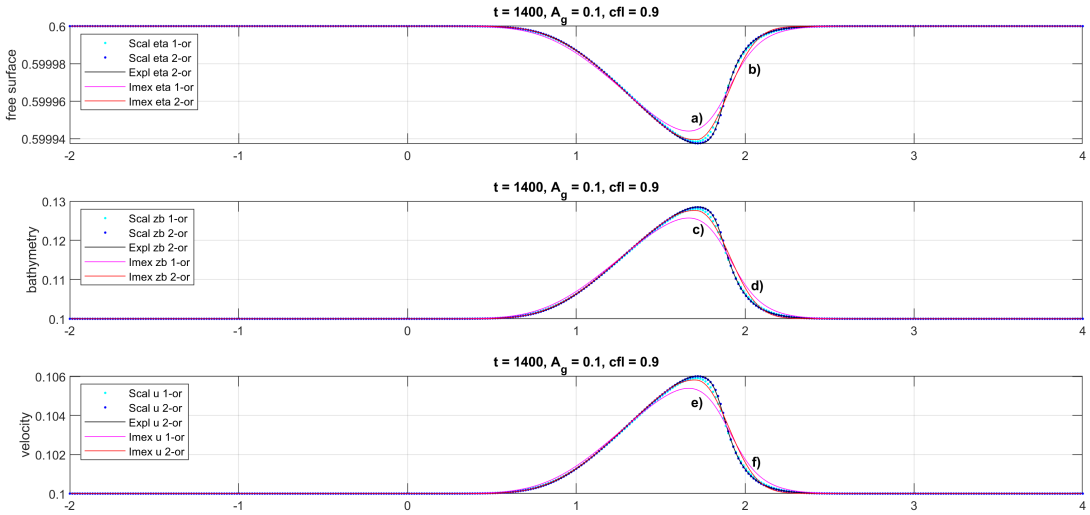

Figure 6.4.2: Test 6.4.1: (1D Exner test). Numerical solutions of thickness (top), sediment layer (center) and velocity (down) for the Exner model on the interval $[-2, 4]$ using a 200−mesh points at time $t = 1400$ with, respectively, $\text{CFL}_{scal} = 0.9$, $\text{CFL}_{expl} = 0.4$ and $\text{CFL}_{IMEX} = 15$.

otherwise, the scalar solutions and the explicit one introduce spurious fluctuation not related with the modeling.

As we expected, there is very good agreement between the solution of the scalar equation and the full system, because the energy associated to fast waves is negligible.

Verified that the semi-implicit strategy leads to results similar to explicit and scalar approximations methods and confirmed that these results, in addition to being similar, are accurate with respect to the expected order, we want to explore the behavior and the results obtained in case a continuous waves group is imposed in the left boundary domain of the

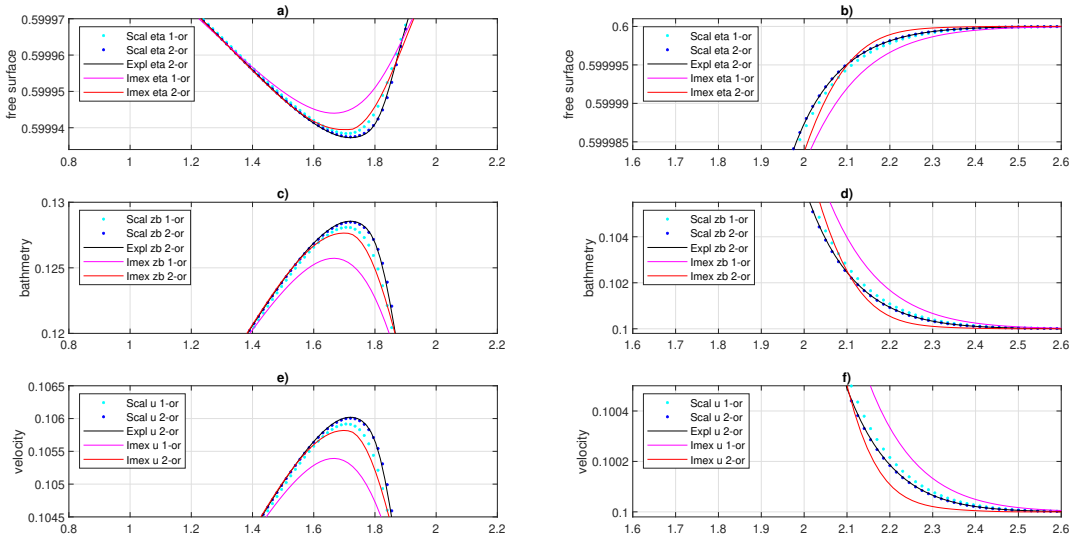

Figure 6.4.3: Test 6.4.1: (1D Exner test). Zoom of critical parts for numerical solutions of thickness (top), sediment layer (center) and velocity (down) for the Exner model at time $t = 1400$ with, respectively, $\text{CFL}_{scal} = 0.9$, $\text{CFL}_{expl} = 0.4$ and $\text{CFL}_{IMEX} = 15$.

h	IMEX or1		IMEX or2		Scal or1		Scal or2		Expl or2	
Points	Order	Error	Order	Error	Order	Error	Order	Error	Order	Error
200	-	4.41E-4	-	5.34E-4	-	1.31E-4	-	3.09E-5	-	8.00E-5
400	0.78	2.58E-4	1.64	1.71E-4	0.91	6.99E-5	1.98	7.85E-6	1.94	2.01E-5
800	0.84	1.44E-4	2.31	3.44E-5	0.96	3.58E-5	2.02	1.94E-6	2.02	5.16E-6
1600	0.90	7.70E-5	2.29	1.83E-6	0.98	1.83E-6	2.01	4.83E-7	2.00	1.29E-6

Table 6.1: Test 6.4.1: (1D Exner test). Errors in L^1-norm and convergence rates related to the sediment z_b for scalar, explicit and semi-implicit scheme at time $t = 1400$ with, respectively, $\text{CFL}_{scal} = 0.9$, $\text{CFL}_{expl} = 0.4$ and $\text{CFL}_{IMEX} = 15$.

velocity u.

6.4.2 1D waves group

Let us consider the one-dimensional Exner system (6.1.8) and the second order semi-implicit method developed before. We want to verify, on the one hand, the temporal evolution of the sediment for very long times, for instance, a final time such that the initial dune has moved 10 times the initial amplitude; on the other hand, whether the presence of fast surface waves under-resolved has a significant effect in the evolution of the initial dune. In this way, we have three different time scales. The slowest one related to the velocity of the dune evolution $3\xi A_g u^2$ due to the Grass equation (6.1.3) if $m = 3$ [50]; the second one related to the water velocity u; the fastest one related to the waves group of order $u + \sqrt{gh}$.

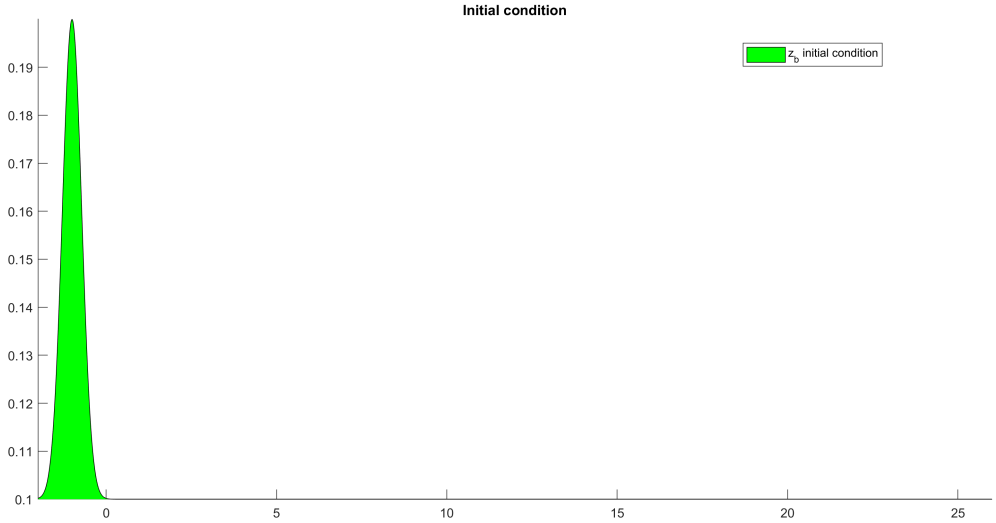

Figure 6.4.4: Test 6.4.1: (1D waves group). Initial condition of sediment for the Exner model on the interval $[-2, 26]$ using a $2000-$mesh points.

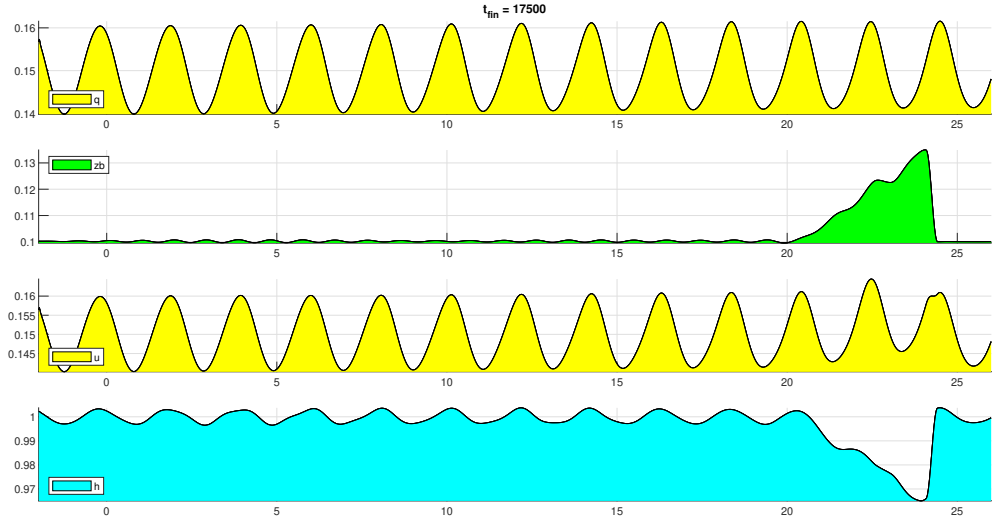

Figure 6.4.5: Test 6.4.1: (1D waves group). Numerical solutions of discharge (top), velocity (center-up), sediment layer (center-down) and thickness (down) for the Exner model on the interval $[-2, 26]$ using a $2000-$mesh points at time $t = 17500$ with CFL$= 9$.

The settings of this test are: $[-2, 26]$ the space domain; $A_g = 0.1$; $\xi = \frac{1}{1-\rho_0}$ where $\rho_0 = 0.2$; $g = 9.81$; $b(x) \equiv 0$; $h_0(-2) = 1$; $u_0(-2) = 0.15$;

$$z_{b_0}(x) = 0.1 + 0.1e^{-\frac{(x+1)^2}{0.4^2}}. \tag{6.4.2}$$

As left side of boundary condition the follow quantities are imposed at ghost points:

$$\begin{bmatrix} h_L \\ q_L \\ z_{bL} \end{bmatrix} = \begin{bmatrix} h(-2) + 1/g(dsig\Delta x + 0.5((u(-2))^2 - sig^2)) \\ sig * h_L \\ z_b(-2) \end{bmatrix}$$

where sig and $dsig$ are respectively the waves group $sig = 0.15 + amp * (sin(frq * t))$ and $dsig = \frac{d\, sig}{dt}$, in which amp and frq are amplitude and frequency of the waves in our case set to 0.01 and 150 respectively.

Figures 6.4.4-6.4.5 show the initial and the numerical solutions for discharge, velocity, sediment layer and thickness obtained with the second-order semi-implicit scheme developed in the previous sections in which the stability condition is set CFL= 9 on the interval $[-2, 26]$ adopting a $2000-$mesh points at time $t = 17500$.

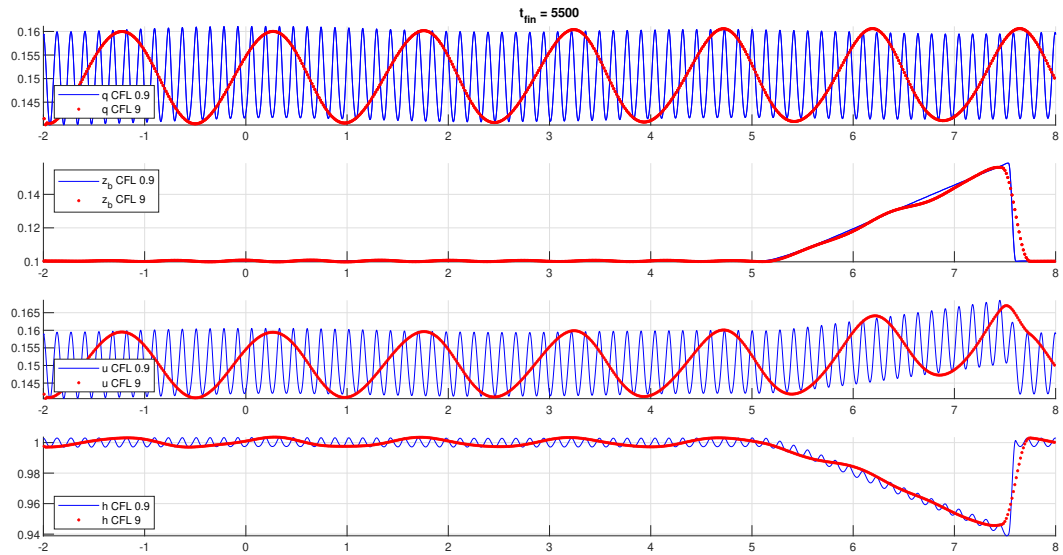

Figure 6.4.6: Test 6.4.1: (1D waves group). Numerical solutions of discharge (top), velocity (center-up), sediment layer (center-down) and thickness (down) for the Exner model on the interval $[-2, 8]$ using a $2000-$mesh points at time $t = 5500$ adopting CFL= 0.9 and CFL= 9.

Since we are imposing a very high-frequency signal to the left part of the domain, the water waves observed are not the real one. In fact, since the period of oscillations is $T = \frac{2\pi}{frq} = 0.042$ and Δt 0.038 (with this settings), the ratio $\frac{T}{\Delta t} = 1.09$ which suggests that, more or less, at each time step a wave is inserted from the signal, so the visible waves on the graph are not the real waves but an understatement of them. To see clearly them a CFL reduction is necessary in order to have more time steps for each wave. Nevertheless, in Figure (6.4.6) we observe that, even if a shock appeared and the free-surface vawes are not resolved, the

semi-implicit method with a low restriction in the stability condition (CFL= 9) is able to capture and properly evolve the sedimentation. Furthermore, we can see how the surface waves group have an active role on the sedimentation but still not affect its evolution. In particular, the solutions obtained resolving the free-surface waves are perfectly in agreement with the solutions obtained with CFL= 9 emphasizing the accuracy of the IMEX strategy.

6.5 2D Exner Model

Let us consider the two-dimensional hyperbolic Shallor water equations

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0 \\ (hu)_t + (hu^2 + \frac{1}{2}gh^2)_x + (hvu)_y = -gh\frac{\partial b}{\partial x} \\ (hv)_t + (huv)_x + (hv^2 + \frac{1}{2}gh^2)_y = -gh\frac{\partial b}{\partial y}, \end{cases} \tag{6.5.1}$$

where (x, y) refers to the Cartesian plane Oxy and t is the time; $h(x, y, t)$, the thickness; $u(x, y, t)$ and $v(x, y, t)$, the horizontal and vertical velocity components; g, the acceleration due to gravity; $b(x, y)$, the bottom topography. In particular, defining the momentum, $m = hu$ and $n = hv$, we get:

$$\begin{cases} h_t + (m)_x + (n)_y = 0 \\ (m)_t + (mu + \frac{1}{2}gh^2)_x + (mv)_y = -ghb_x \\ (n)_t + (nu)_x + (nv + \frac{1}{2}gh^2)_y = -ghb_y. \end{cases} \tag{6.5.2}$$

The system of equations used in this section is obtained by coupling 2D shallow water equation (6.5.2) and the 2D sediment equation:

$$(z_b)_t + (q_{x,b})_x + (q_{y,b})_y = 0 \tag{6.5.3}$$

where $z_b(x, y, t)$ represents the height of the sediment layer and, $q_{x,b}(u, v)(x, y, t)$ and $q_{y,b}(u, v)(x, y, t)$, the solid transport discharge parameters, in our case computed by the 2D Grass model

[50, 63, 94]

$$q_{x,b} = \xi A_g u(u^2+v^2)^{\frac{m-1}{2}} \tag{6.5.4}$$

$$q_{y,b} = \xi A_g v(u^2+v^2)^{\frac{m-1}{2}}, \tag{6.5.5}$$

with $m \in [1,4] \cap \mathbb{N}$, $A_g \in]0,1[$ and $\xi = \dfrac{1}{1-\rho_0}$ where ρ_0 is the porosity of the sediment layer.

In this way, the 2D Exner system of balance laws is so get:

$$\begin{cases} h_t + (m)_x + (n)_y = 0 \\ (m)_t + (mu + \frac{1}{2}gh^2)_x + (mv)_y = -gh(b+z_b)_x \\ (n)_t + (nu)_x + (nv + \frac{1}{2}gh^2)_y = -gh(b+z_b)_y \\ (z_b)_t + (q_{x,b})_x + (q_{y,b})_y = 0. \end{cases} \tag{6.5.6}$$

Observe that, assuming flat bottom topography $b(x,y) \equiv 0$, the system (6.5.6) could be written in the following way:

$$\partial_t U + A_1(U)\partial_x U + A_2 \partial_y U = 0$$

where

$$U = \begin{bmatrix} h \\ m \\ n \\ z_b \end{bmatrix}; \quad A_1(U) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ gh-u^2 & 2u & 0 & gh \\ -uv & v & u & 0 \\ \alpha_x & \beta_x & \gamma_x & 0 \end{bmatrix}; \quad A_2(U) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ gh-v^2 & 0 & 2v & gh \\ \alpha_y & \beta_y & \gamma_y & 0 \end{bmatrix};$$

in which the terms α_s, β_s and γ_s, with $s \in \{x,y\}$, represent the $q_{s,b}$ derivatives respect to h, m and n, i.e. for $s \in \{x,y\}$

$$\alpha_s = \frac{\partial q_{s,b}}{\partial h}, \quad \beta_s = \frac{\partial q_{s,b}}{\partial m}, \quad \gamma_s = \frac{\partial q_{s,b}}{\partial n}.$$

System (6.5.6) is hyperbolic if and only if the characteristic polynomials related to $A_1(U)$

and $A_2(U)$:

$$p_\lambda(\lambda) = (u-\lambda)\Big[-\lambda\Big((u-\lambda)^2 - gh\Big) + gh\Big(\beta_x\lambda + \alpha_x + \gamma_x v\Big)\Big]$$
$$p_\mu(\mu) = (v-\mu)\Big[-\mu\Big((v-\mu)^2 - gh\Big) + gh\Big(\gamma_y\mu + \alpha_y + \beta_y u\Big)\Big]$$

have four distinct root, assuming $\lambda_4 = u$ and $\mu_4 = v$, such as $\lambda_1 < \lambda_2 < \lambda_3$ and $\mu_1 < \mu_2 < \mu_3$ respectively. In our case, assuming that the interaction between the water and the sediment is weak, we are looking for a numerical scheme that are able to capture accurately the evolution of the sediment when the interactions are small. For this reason, we suppose that the corresponding numerical viscosity terms for the LLF fluxes in the x and y direction are the eigenvalues closer to zero, i.e. λ_2 and μ_2. In order to compute these eigenvalues, we use an iterative root finding algorithm, such as Newton method etc., and used the numerical viscosity parameter for the corresponding sediment transport equation (6.5.4)-(6.5.5).

At the end, let us rewrite the 2D Exner system (6.5.6) in function of η where $\eta(x,y,t) = h(x,y,t) + b(x,y) + z_b(x,y,t)$ represents the elevation of the undisturbed water surface. In practise, system (6.5.6) becomes:

$$\begin{cases} \eta_t + (m + q_{x,b})_x + (n + q_{y,b})_y = 0 \\ (m)_t + (mu)_x + (mv)_y + gh(\eta)_x = 0 \\ (n)_t + (nu)_x + (nv)_y + gh(\eta)_y = 0 \\ (z_b)_t + (q_{x,b})_x + (q_{y,b})_y = 0. \end{cases} \tag{6.5.7}$$

6.6 2D semi implicit scheme

In this section, we will present the extension of the 1D semi implicit scheme for the non-conservative Exner model (6.1.8) to the 2D Exner model (6.5.7). We will aim at an implicit treatment of the surface water waves while the corresponding slow sediment wave is treated explicitly. As we have done in Section 6.2, we will just emphasize the first and second order reconstruction in space and time.

Let us consider a partition of the rectangular $[a_1, b_1] \times [a_2, b_2]$ in cells defined by $I_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, with $i, j \in \mathbb{N}$. For sake of simplicity, we adopt a uniform Cartesian mesh direction by direction with mesh spacing, respectively, Δx and Δy, i.e. $x_i = a_1 + (i -$

$\frac{1}{2})\Delta x$ and $y_j = a_2 + (j - \frac{1}{2})\Delta y$. As previously, Δt is the time step such that $t^n = n\Delta t$.

Finally, we denote by $U^n_{i,j}$ an approximation on the mean value of U over cell $I_{i,j}$ at time $t = t^n$ as:

$$U^n_{i,j} \cong \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} U(x, y, t^n) dy dx.$$

6.6.1 First order scheme

Let us consider the system in non-conservative form (6.5.7). A semi-discrete in time first order semi implicit method could be written as:

$$\begin{cases} \eta^{n+1} = \eta^n - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b}) - \Delta t D_x(m^{n+1}) - \Delta t D_y(n^{n+1}), \\ m^{n+1} = m^n - \Delta t \hat{D}_x(m^n u^n) - \Delta t \hat{D}_y(m^n v^n) - \Delta t g h^n D_x(\eta^{n+1}), \\ n^{n+1} = n^n - \Delta t \hat{D}_x(n^n u^n) - \Delta t \hat{D}_y(n^n v^n) - \Delta t g h^n D_y(\eta^{n+1}), \\ z_b^{n+1} = z_b^n - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b}), \end{cases} \tag{6.6.1}$$

where the differential operators $D_x, D_y, \hat{D}_x$ and $\hat{D}_y$ are defined as in Section 6.2 direction by direction.

For the sake of simplicity, let us rewrite system (6.5.7) into (6.6.1) to emphasize the explicit from implicit part in the following way:

$$\begin{cases} m^* = m^n - \Delta t \hat{D}_x(m^n u^n) - \Delta t \hat{D}_y(m^n v^n); \\ n^* = n^n - \Delta t \hat{D}_x(n^n u^n) - \Delta t \hat{D}_y(n^n v^n); \\ \eta^* = \eta - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b}) - \Delta t \hat{D}_x(m^*) - \Delta t \hat{D}_y(n^*); \\ \eta^{n+1} = \eta^* + g\Delta t^2 D_x(h^n D_x(\eta^{n+1})) + g\Delta t^2 D_y(h^n D_y(\eta^{n+1})); \\ m^{n+1} = m^* - g\Delta t h^n D_x(\eta^{n+1}); \\ n^{n+1} = n^* - g\Delta t h^n D_y(\eta^{n+1}); \\ z_b^{n+1} = z_b^n - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b}). \end{cases} \tag{6.6.2}$$

The procedure to solve system (6.6.1), hence (6.6.2), is:

1. solve explicitly $m^* = m^n - \Delta t \hat{D}_x(m^n u^n) - \Delta t \hat{D}_y(m^n v^n)$ as

$$m^*_{i,j} = m^n_{i,j} - \frac{\Delta t}{\Delta x}\Big(F^1_{i+\frac{1}{2},j} - F^1_{i-\frac{1}{2},j}\Big) - \frac{\Delta t}{\Delta y}\Big(G^1_{i,j+\frac{1}{2}} - G^1_{i,j-\frac{1}{2}}\Big),$$

where F^1 and G^1 are computed direction by direction with the corresponding Rusanov reconstruction defined on Section 6.2;

2. solve explicitly $n^* = n^n - \Delta t \hat{D}_x(n^n u^n) - \Delta t \hat{D}_y(n^n v^n)$ as

$$n^*_{i,j} = n^n_{i,j} - \frac{\Delta t}{\Delta x}\Big(F^2_{i+\frac{1}{2},j} - F^2_{i-\frac{1}{2},j}\Big) - \frac{\Delta t}{\Delta y}\Big(G^2_{i,j+\frac{1}{2}} - G^2_{i,j-\frac{1}{2}}\Big),$$

where F^2 and G^2 are computed direction by direction with the Rusanov flux;

3. solve explicitly $\eta^* = \eta - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b}) - \Delta t \hat{D}_x(m^*) - \Delta t \hat{D}_y(n^*)$ as

$$\begin{aligned}\eta^*_{i,j} = \eta_{i,j} &- \frac{\Delta t}{\Delta x}\Big(F^3_{i+\frac{1}{2},j} - F^3_{i-\frac{1}{2},j}\Big) - \frac{\Delta t}{\Delta y}\Big(G^3_{i+\frac{1}{2},j} - G^3_{i-\frac{1}{2},j}\Big) + \\ &- \frac{\Delta t}{\Delta x}\Big(F^4_{i+\frac{1}{2},j} - F^4_{i-\frac{1}{2},j}\Big) - \frac{\Delta t}{\Delta y}\Big(G^4_{i+\frac{1}{2},j} - G^4_{i-\frac{1}{2},j}\Big),\end{aligned}$$

where F^3 and G^3 are the Rusanov operators referred to $q_{x,b}$ and $q_{y,b}$; while F^4 and G^4 are related to m^* and n^*;

4. fixed $k_x = g(\frac{\Delta t}{\Delta x})^2$ and $k_y = g(\frac{\Delta t}{\Delta y})^2$, solve implicitly $\eta^{n+1} = \eta^* + g\Delta t^2 D_x(h^n D_x(\eta^{n+1})) + g\Delta t^2 D_y(h^n D_y(\eta^{n+1}))$ as

$$\begin{aligned}&\eta^{n+1}_{i,j}\Big(1 - k_x(h_{i,j-\frac{1}{2}} + h_{i,j+\frac{1}{2}}) - k_y(h_{i-\frac{1}{2},j} + h_{i+\frac{1}{2},j})\Big) + \\ &+\eta^{n+1}_{i,j-1}\Big(k_x(h_{i,j-\frac{1}{2}}\Big) + \eta^{n+1}_{i,j+1}\Big(k_y(h_{i,j+\frac{1}{2}}\Big) + \\ &+\eta^{n+1}_{i-1,j}\Big(k_x(h_{i-\frac{1}{2},j}\Big) + \eta^{n+1}_{i+1,j}\Big(k_y(h_{i+\frac{1}{2},j}\Big) = \eta^*_{i,j}.\end{aligned}$$

This is an invertible linear system which can be solved to detect $\eta^{n+1} = [\eta^{n+1}_{i,j}]$ for all $i = 1, \dots, N_x$ and $j = 1, \dots, N_y$;

5. solve explicitly $m^{n+1} = m^* - g\Delta t h^n D_x(\eta^{n+1})$ as

$$m^{n+1}_{i,j} = m^*_{i,j} - g\frac{\Delta t}{\Delta x} h^n_{i,j}\Big(\eta^{n+1}_{i+\frac{1}{2},j} - \eta^{n+1}_{i-\frac{1}{2},j}\Big),$$

where $\eta^{n+1}_{i\pm\frac{1}{2},j} = \frac{1}{2}\Big(\eta^{n+1}_{i,j} + \eta^{n+1}_{i\pm1,j}\Big)$ for all $j = 1, \dots, N_y$;

6. solve explicitly $n^{n+1} = n^* - g\Delta t h^n D_y(\eta^{n+1})$ as

$$n^{n+1}_{i,j} = n^*_{i,j} - g\frac{\Delta t}{\Delta y} h^n_{i,j}\Big(\eta^{n+1}_{i,j+\frac{1}{2}} - \eta^{n+1}_{i,j-\frac{1}{2}}\Big),$$

where $\eta^{n+1}_{i,j\pm\frac{1}{2}} = \frac{1}{2}\Big(\eta^{n+1}_{i,j} + \eta^{n+1}_{i,j\pm1}\Big)$ for all $i = 1, \dots, N_x$;

7. solve explicitly $z^{n+1}_b = z^n_b - \Delta t \hat{D}_x(q^n_{x,b}) - \Delta t \hat{D}_y(q^n_{y,b})$ as

$$z^{n+1}_{b_{i,j}} = z^{n+1}_{b_{i,j}} - \frac{\Delta t}{\Delta x}\Big(F^5_{i+\frac{1}{2},j} - F^5_{i-\frac{1}{2},j}\Big) - \frac{\Delta t}{\Delta y}\Big(G^5_{i+\frac{1}{2},j} - G^5_{i-\frac{1}{2},j}\Big),$$

in which F^5 and G^5 are computed direction by direction with the Rusanov reconstruction;

8. compute $h^{n+1}_{i,j} = \eta^{n+1}_{i,j} - b_{i,j} - z^{n+1}_{b_{i,j}}$, for all $i = 1, \dots, N_x$ and for all $j = 1, \dots, N_y$.

6.6.2 Second order scheme

Let us consider the system in non-conservative form (6.5.7), the second order reconstruction in time is obtained with a 2D IMEX second order Runge-Kutta method [7, 8]. For this reason, let us rewrite the system (6.5.7) in the partitioned ODE form in which the first component is treated explicitly and the second component implicitly. The ODE system is then so defined:

$$U' = H(U, U), \tag{6.6.3}$$

where, in the 2D case, $U = [\eta, m, n, z_b]^T$ and $H(U, U)$ is defined as:

$$H(U,U) = \begin{bmatrix} -(q_{x,b} + m)_x - (q_{y,b} + n)_y \\ -(mu)_x - gh(\eta)_x - (mv)_y \\ -(nu)_x - (nv)_y - gh(\eta)_y \\ -(q_{x,b})_x - (q_{y,b})_y \end{bmatrix} \tag{6.6.4}$$

that, differentiating between explici and implicit part, the partitioned system is:

$$H(U_E, U_I) = \begin{bmatrix} -\hat{D}_x((q_{x,b})_E) - \hat{D}_y((q_{y,b})_E) & -D_x(m_I) - D_y(n_I) \\ -\hat{D}_x((mu)_E) - \hat{D}_y((mv)_E) & -gh_E D_x(\eta_I) \\ -\hat{D}_x((nu)_E) - \hat{D}_y((nv)_E) & -gh_E D_y(\eta_I) \\ -\hat{D}_x((q_{x,b})_E) - \hat{D}_y((q_{y,b})_E) & \end{bmatrix}. \tag{6.6.5}$$

Following the same reconstruction used for the 1D model, the procedure to update the numerical solution for (6.6.3) is:

1. $U^{(1)}_E = U^n$;

2. $U_I^{(1)} = U^n + \Delta t \gamma H(U_E^{(1)}, U_I^{(1)})$;

3. $U_E^{(2)} = (1 - \frac{c}{\gamma})U^n + \frac{c}{\gamma}U_I^{(1)}$;

4. $U_I^{(2)} = (1 - \frac{1-\gamma}{\gamma})U^n + \frac{1-\gamma}{\gamma}U_I^{(1)} + \Delta t \gamma H(U_E^{(2)}, U_I^{(2)})$;

5. $U^{n+1} = U_I^{(2)}$.

6.7 2D Exner numerical experiments

In this section we check the semi-implicit scheme for the 2D Exner model with two different initial conditions: a parabolic and a conical sediment.

6.7.1 Parabolic Sediment

With this purpose in mind, let us consider the 2D Exner model (6.5.7) where initial conditions are so set: $\eta_0(x, y, 0) = 0.6$, $b(x, y) = 0$, $m(x, y) = 0.1$, $n(x, y) = 0.01$ and

$$z_{b_0}(x, y) = 0.1 + 0.006e^{-\frac{(x-0.4)^2}{0.4^2}} \tag{6.7.1}$$

a one-dimensional parabolic sediment. Free boundary conditions are imposed at ghost points, see Figure 6.7.1. The numerical results are obtained with the second order semi-implicit

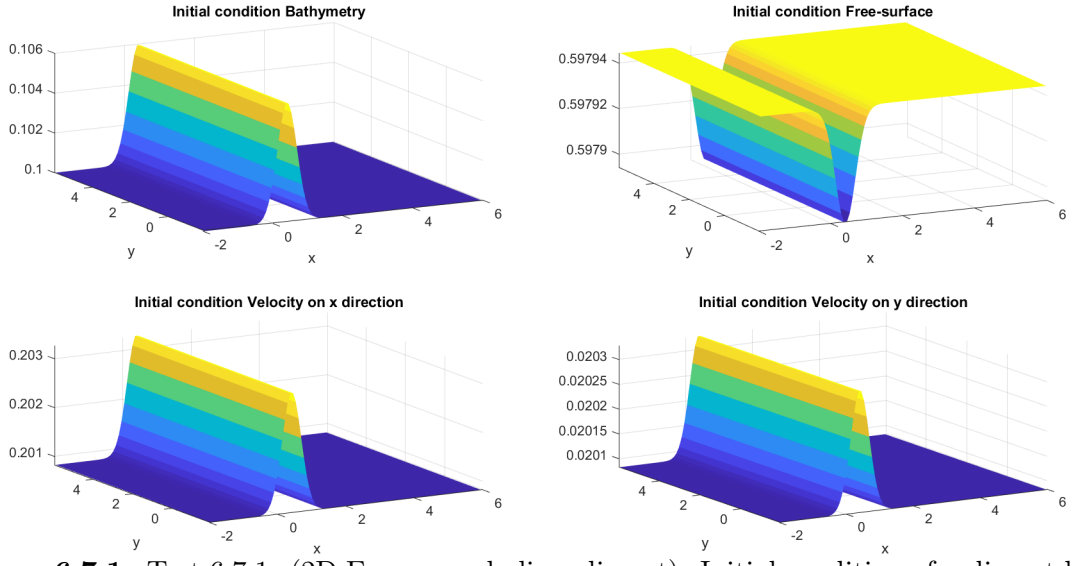

Figure 6.7.1: Test 6.7.1: (2D Exner parabolic sediment). Initial condition of sediment layer (top-left), thickness (top-right) and velocity (down) for the 2D Exner model on the square $[-2, 6]\times[-2, 6]$ using a 100×100 mesh points.

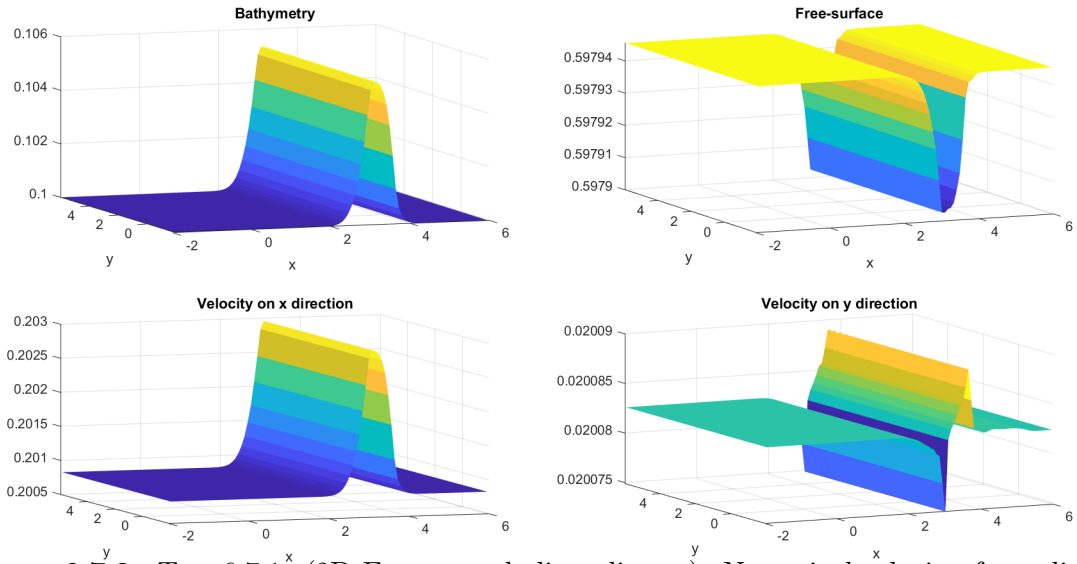

Figure 6.7.2: Test 6.7.1: (2D Exner parabolic sediment). Numerical solution for sediment layer (top-left), thickness (top-right) and velocity (down) for the 2D Exner model on the square $[-2,6]\times[-2,6]$ using a 100×100 mesh points at time $t=450$ with CFL$=6$.

scheme introduced on Section 6.6.2 on the square $[-2,6]\times[-2,6]$ adopting a 100×100 mesh points, CFL$=9$ at time $t=450$. As it shown in Figure 6.7.2, the numerical results are in accordance with the one-dimensional one.

6.7.2 Conical Sediment

As last experiment we consider a fully two-dimensional conical sediment. For this reason, let us consider the 2D Exner model (6.5.7) where initial conditions are so set: $\eta_0(x,y,0)=0.6$, $b(x,y)=0$, $m(x,y)=0.1$, $n(x,y)=0$ and

$$z_{b_0}(x,y)=0.1+0.006e^{-\frac{(x-0.4)^2}{0.4^2}-(y-3)^2} \tag{6.7.2}$$

a conical sediment, see Figure 6.7.3. Free boundary conditions are imposed at ghost points.

The numerical results are obtained with the second order semi-implicit scheme introduced on Section 6.6.2 on the square $[-2,6]\times[-2,6]$ adopting a 100×100 mesh points, CFL$=6$ at time $t=450$. Figure 6.7.4 shows the numerical results for bathymetry, free-surface and velocity in both directions. Particular attention was paid to sediment evolution in Figure 6.7.5. The second order semi-implicit scheme is able to perform the sedimental evolution as expected.

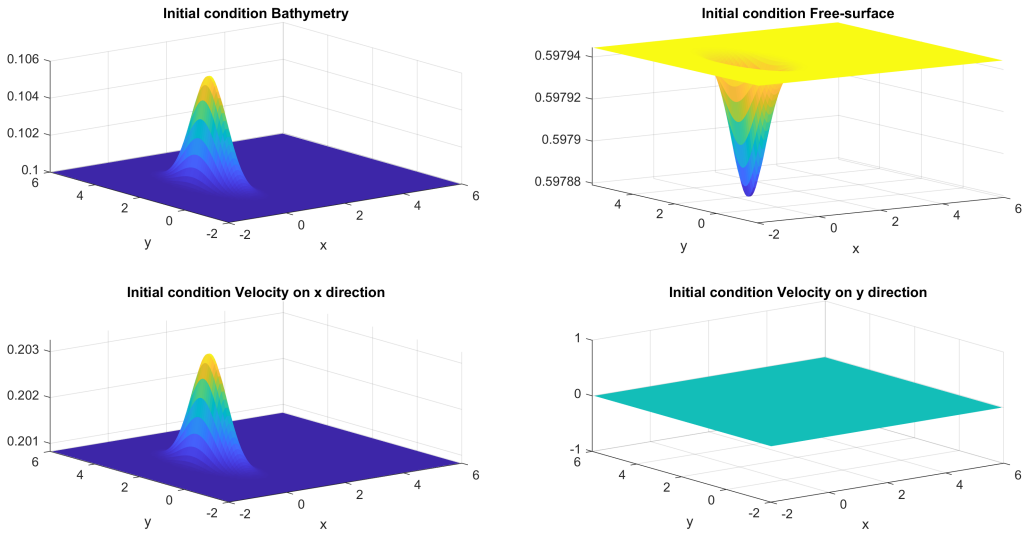

Figure 6.7.3: Test 6.7.2: (2D Exner conical sediment). Initial condition of sediment layer (top-left), thickness (top-right) and velocity (down) for the 2D Exner model on the square $[-2,6]\times[-2,6]$ using a 100×100 mesh points.

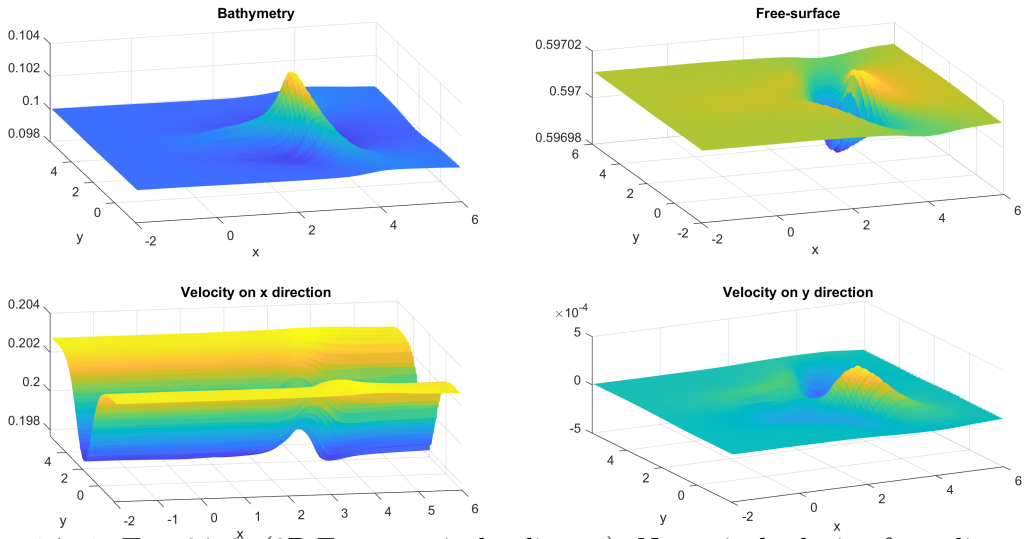

Figure 6.7.4: Test 6.7.2: (2D Exner conical sediment). Numerical solution for sediment layer (top-left), thickness (top-right) and velocity (down) for the 2D Exner model on the square $[-2,6]\times[-2,6]$ using a 100×100 mesh points at time $t = 450$ with CFL$= 6$.

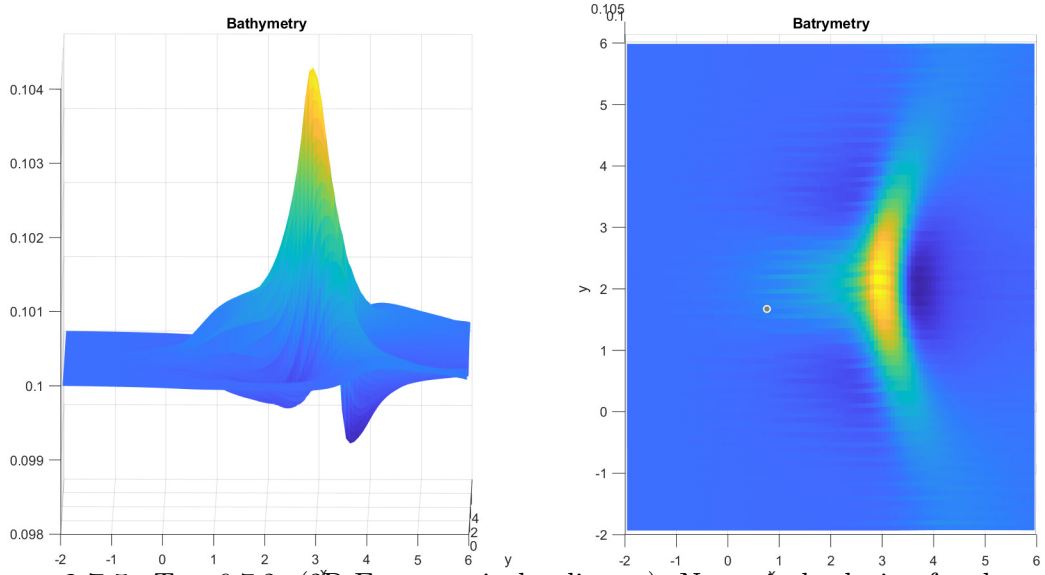

Figure 6.7.5: Test 6.7.2: (2D Exner conical sediment). Numerical solution for the sediment layer on the square $[-2, 6] \times [-2, 6]$ using a 100×100 mesh points at time $t = 450$ with CFL$= 6$.

Chapter 7

Conclusion

This work deals mainly with the construction, analysis, implementation and testing of high-order shock-capturing Adaptive Compact Approximate Taylor (ACAT) methods to the treatment of hyperbolic systems of conservation and balance laws and their well-balanced version. These methods were previously developed as an order adaptive version of the Compact Approximate Taylor Methods for hyperbolic systems of conservation laws, introduced in [14], in which the solution at every point is updated using the stencil of maximal length for which the solution is smooth [15]. Successively, an extension to systems of balance laws has been introduced [13]. The starting point for systems of balance laws is to rewrite the systems as conservation laws, by subtracting to the flux a primitive of the the source term. Meanwhile, the well-balanced property has been obtained rewriting the systems as conservation laws, by subtracting to the standard flux the one corresponding to the stationary solution, and adding to the primitive of the source the source computed at the stationary solution. The methods are developed for systems in one and two space dimensions, and could be extended to 3D. In principle the procedure allows the construction of well-balanced schemes of arbitrary order, although the computational complexity quickly increases with the order of accuracy. We prove that the constructed schemes are exactly fully well-balanced because they are able to preserve any stationary solutions.

Concerning the one-dimensional and two-dimensional systems of conservation laws (*Chapter 3*), the heuristic analysis and the details of Compact Approximate Taylor methods and the corresponding order-adaptive technique to avoid the spurious oscillations close the discontinuities were presented. Several numerical results, obtained with the new family of methods (ACAT), have been compared with the corresponding WENO-RK methods (Finite

Differences WENO reconstructions in space, TVD-RK in time). The linear transport equation, Burgers equation, the 1D and 2D compressible Euler equations have been considered. For CFL ≤ 0.5 all the numerical solutions work as expected, and the results obtained with WENO or ACAT methods are similar. For CFL$\in$ [0.5) WENO schemes may introduce oscillations, while the ACAT is generally oscillation-free. The possibility of using larger CFL condition and consequently larger time steps, compensate the extra computational cost of ACAT. When the solution is not sufficiently regular, or when high order accuracy is not required, the most cost-effective ACAT scheme is the second order one.

Concerning the one-dimensional (*Chapter 4*) and two-dimensional (*Chapter 5*) systems of balance laws, the developments and the details of well-balanced and non well-balanced CAT schemes and its order-adaptive strategy to avoid the introduction of spurious oscillations were presented. A set of numerical results obtained with the well-balanced and non well-balanced schemes have been compared with exact or reference solutions. The linear transport equation with source, the Burgers equation with source, the 1D Shallow-Water equations, the 1D and 2D compressible Euler equations with gravity have been considered. The use of suitable limiters allow an effective treatment of discontinuous solutions. In all cases we observe that stationary solutions are preserved within machine precision, allowing very accurate results when the solution is a small deviation from equilibrium or an initial condition far from the stationary solution is considered. The numerical solutions and the errors are in accordance with the expected order, nevertheless the well-balanced schemes are able to preserve the stationary solutions with machine precision and return better solutions compared with the non well-balanced methods when a small perturbation of the stationary solution has been considered as initial condition. Some order reduction phenomena have been observed more frequently in the non well-balanced reconstructions to be associated mainly with the family of smoothness indicators examined to capture the regularity of the numerical solutions.

The main advantage of the ACAT method for systems of conservation and balance laws consists in its generality: it allows the automatic construction of very high order well-balanced schemes for multi-dimensional systems. The main disadvantage is the extra computational cost due to the evaluations of the Taylor expansion terms. This drawback could be alleviated by parallel implementation that would have a significant decrease in the computational cost because of the local nature of the method.

The last part of the thesis (*Chapter 6*) concerns a *work in progress* on the development of semi-implicit schemes for the 1D and 2D Exner model of shallow water with sedimentation.

The objective was to drastically improve the efficiency in the computation of the evolution of the sediment by treating water waves implicitly, thus allowing much larger time steps than the one permotted by standard CFL condition on explicit schemes. This procedure has a theoretical basis in the following wording. After some manipulation, the Exner model can be written as a non-conservative hyperbolic system

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = 0.$$

This system is strictly hyperbolic if and only if the characteristic polynomial has three distinct real roots $\lambda_1 < \lambda_2 < \lambda_3$. Under the hypothesis of Froude number (Fr) less then 1 it is $\lambda_1 < 0$ and $\lambda_3 > 0$. Assuming that the interaction between the water and the sediment is weak, it is $\lambda_2 \leq \min(|\lambda_1|, |\lambda_3|)$, i.e. the wave speed of the sediment is much smaller than the water wave speeds. An explicit method implies a strong stability restriction due to the velocity of the free-surface wave. This restriction involves in a very long computation time that could be reduced neglecting the behaviour of the free-surface waves behaviour and looking at the sediment evolution. We want to check that even if we do not resolve the small time scale of the waves, still the semi-implicit method is able to correctly capture the sediment evolution. To this purpose, a simplified model in which the flow is quasi-stationary has been considered. As expected, there is very good agreement between the solution of the scalar equation and the full system, because in this case the energy associated to fast waves is negligible. Successively,the long-term behaviour of the sediment must be checked even in the presence of under-resolved fast water waves and, if necessary, the effects of these on the sediment must be analysed. This exploration is subject of current investigation.

There are still a few things that require improvement and generalization:

- Optimal implementation of CAT methods in GPU architectures. The implementation of ACAT or WBACAT methods for systems of conservation and balance laws carried out to compute the numerical results shown in this work is not optimal and does not take advantage of the potentiality of these methods: they are highly parallelizable and do not need the storage of intermediate temporal stages. Therefore, the comparisons of computational costs or efficiency curves shown in the previous chapters lead only to partial conclusions. Next developments include the implementation of the methods in GPU architectures and the systematic comparison between them.

- Combination of CAT methods with a new adaptive non a priori strategy. In fact, instead of using a priori smoothness indicators to cure the spurious oscillations close to discontinuities, a posteriori analysis of the updated numerical solution could be used such as MOOD approach see [26, 27, 89]. This analysis is performed at every time step and it is followed by a local recalculation of the solution where it is necessary using a more robust numerical method. Besides the spurious oscillations, this methodology allows one to control aspects such as the positivity of the numerical method. CAT methods are excellent candidates to be combined with this technique, due to their good stability properties and the minimal size of their stencils. The idea would be to update the numerical solutions at every time step with CAT2P. Then, this first numerical solution is analyzed and the cells were wrong solutions are detected are marked. Next, the numerical solutions at the marked cells are computed again using now CAT2($P-1$). This new numerical solution is then analyzed and the procedure follows in a recursive way. In the worst-case scenario, the numerical solution will be updated in part of the domain with a robust first order numerical method. This strategy may lead to efficient and robust high-order numerical methods.

- ACAT-IMEX coupling. An open problem is how to couple ACAT methodology with implicit-explicit method for the treatment of problems with stiff source. In order to obtain, on the one hand, a computationally more expensive method that is able to solve problems with stiff source without introducing limitations on the CFL stability condition.

- More realistic Exner models. Other equations for the evolution of the sediment can be considered making the simulations more sophisticated and obtaining numerical results that are more consistent with the experimental data.

- Semi-Implicit schemes for multi-layer shallow water. A multi-layer semi-implicit approach could be explored to couple an implicit treatment of the free-surface waves in the top layer and an explicit treatment for the sediment evolution in the bottom layer.

Bibliography

[1] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable wellbalanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25(6):2050–2065, 2004.

[2] A. Baeza, R. Bürger, P. Mulet, and D. Zorío. On the efficient computation of smoothness indicators for a class of WENO reconstructions. *Journal of Scientific Computing*, 80:1240–1263, 2019.

[3] A. Baeza, R. Bürger, P. Mulet, and D. Zorío. An efficient third-order WENO scheme with unconditionally optimal accuracy. *SIAM Journal on Scientific Computing (To appear)*, 2020.

[4] J. P. Berberich, R. Käppeli, P. Chandrashekar, and C. Klingenberg. High order discretely well-balanced methods for arbitrary hydrostatic atmospheres. *Communications in Computational Physics*, 30(3):666–708, 2021.

[5] A. Bermúdez and M. E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1994., 1049-1071.

[6] S. Bertoluzza, S. Falletta, G. Russo, and C. Shu. *Numerical Solutions of Partial Differential Equations*. Advanced Courses in Mathematics CRM Barcelona Birkhauser, 1 edition, 2009.

[7] S. Boscarino. Error analysis of IMEX Runge–Kutta methods derived from Differential-Algebraic systems. *SIAM J. Numer. Anal.*, 45(4):1600–1621, 2006.

[8] S. Boscarino, F. Filbet, and G. Russo. High order semi-implicit schemes for time dependent partial differential equations. *Journal of Scientific Computing*, 68(8):975–1001, 2016.

[9] S. Boscarino, L. Pareschi, and G. Russo. A unified imex runge-kutta approach for hyperbolic systems with multiscale relaxation. *SIAM J. Numer. Anal.*, 55(4):2017, 2085-2109.

[10] F. Bouchut. Non-linear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources. *Frontiers in Mathematics*, Birkhauser, 2004.

[11] A. Bressan. *Hyperbolic systems of conservation laws: the one-dimensional Cauchy problem.* Oxford University Press, 2000.

[12] H. Brezis and F. Browder. Partial Differential Equations in the 20th Century. *Advances in Mathematics*, 135(1):76–144, 1998.

[13] H. Carrillo, E. Macca, C. Parés, and G. Russo. An order-adaptive compact approximate taylor method for systems of balance law and relative well-balanced scheme. *Journal of Computational Physics, Submitted*, 2022.

[14] H. Carrillo, E. Macca, C. Parés, G. Russo, and D. Zorío. An order-adaptive compact approximate taylor method for systems of conservation law. *Journal of Computational Physics*, 438:31, 2021.

[15] H. Carrillo and C. Parés. Compact approximate taylor methods for systems of conservation laws. *J. Sci. Comput.*, 80:1832–1866, 2019.

[16] V. Caselles, R. Donat, and G. Haro. Flux-gradient and source-term balancing for certain high resolution shock-capturing schemes. *Computers & Fluids*, 38:2009., 16-36.

[17] M. Castro, C. Chalons, and T. M. de Luna. A fully well-balanced lagrange-projection-type scheme for the shallow-water equations. *SIAM J. Numer. Anal.*, 56(6):3071–3098, 2018.

[18] M. Castro, E. D. Fernández-Nieto, and A. M. Ferreiro. Sediment transport models in shallow water equations and numerical approach by high order finite volume methods. *Comput. & Fluids*, 37(3):299–316, 2008.

[19] M. Castro, I. Gómez-Bueno, and C. Parés. High-order well-balanced methods for systems of balance laws: a control-based approach. *Submitted*, 2020.

[20] M. Castro, J. López-García, and C. Parés. High order exactly well-balanced numerical methods for shallow water systems. *Journal of Computational Physics*, 246:242–264, 2013.

[21] M. Castro and C. Parés. Well-balanced high-order finite volume methods for systems of balance laws. *J. Sci. Comput.*, 82:48, 2020.

[22] V. Casulli and E. Cattani. Stability, accuracy and efficiency of a semi-implicit method for three-dimensional shallow water flow. *Comput. Math. Appl.*, 27(4):99–112, 1994.

[23] P. Chandrashekar and C. Klingenberg. A second order well-balanced finite volume scheme for Euler equations with gravity. *SIAM J. Sci. Comput.*, 37:383–402, 2015.

[24] A. Chertock, S. Cui, A. Kurganov, S. Özcan, and E. Tadmor. Well-balanced schemes for the Euler equations with gravitation: Conservative formulation using global fluxes. *J. Comp. Phys.*, 358:36–52, 2018.

[25] A. J. Chorin and J. E. Marsden. *A mathematical introduction of fluid mechanics*. Springer, 3 edition, 2000.

[26] S. Clain, S. Diot, and R. Loubère. A high-order finite volume method for systems of conservation laws: Multi-dimensional Optimal Order Detection (mood). *Journal of computational Physics*, 230(10):4028–4050, 2011.

[27] S. Clain, S. Diot, and R. Loubère. Multi-dimensional Optimal Order Detection (MOOD): a very high-order finite volume scheme for conservation laws on unstructured meshes. *Finite Volumes for Complex Applications VI Problems & Perspectives*, 6:263.271, 2011.

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

[29] R. Courant and D. Hilbert. *Methods of Mathematical Physics vol II*. New York: Wiley-Interscience, 1962.

[30] C. M. Dafermos. Polygonal approximations of solutions of the initial value problem for a conservation law. *J. Math. Anal. Appl.*, 38:33–41, 1972.

[31] G. De Marco. *Analisi Matematica Due/1*. Decibel-Zanichelli, 2004.

[32] G. De Marco. *Analisi Matematica Due/2.* Decibel-Zanichelli, 2005.

[33] R. Donat and A. Marquina. Capturing shock reflections: an improved flux formula. *J. Comput. Phys.*, 125:42–58, 1996.

[34] R. Donat and A. Martínez-Gavera. Hybrid second order schemes for scalar balance laws. *Journal of Scientific Computing*, 48:52–69., 2011.

[35] M. Dumbser, D. Balsara, E. Toro, and C. Munz. A unified framework for the construction of one-step finite-volume and discontinuous galerkin schemes. *Journal of Computational Physics*, 227:8209–8253, 2008.

[36] B. Einfeldt, P. Roe, C. Munz, and B. Sjogreen. On Godunov–type methods near low densities. *Journal of Computational Physics*, 92:273–295, feb 1991.

[37] C. Enaux, M. Dumbser, and E. Toro. Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. *Journal of Computational Physics*, 227(2):3971–4001, 2008.

[38] L. C. Evans. *Partial differential equations, Volume 19 of Graduate Studies in Mathematichs.* American Mathematical Society, Providence, RI, 1998.

[39] C. F. Faà di Bruno. Note sur un nouvelle formule de calculo différentiel. *Quart. J. Math.*, 1:359–360, 1857.

[40] G. B. Folland. *Partial Differential Equation.* Princeton University Press, 2 edition, 1995.

[41] B. Fornberg. Generation of finite difference formulas on arbitrarily space grids. *Mathematics of Computation*, 51:699–706., 1988.

[42] B. Fornberg. Classroom note: calculation of weights in finite difference formulas. *SIAM Review*, 40:685–691., 1998.

[43] W. Fulton and J. Harris. *Representation theory. A first course. Graduate Texts in Mathematics, Readings in Mathematics.* New York: Springer-Verlag, 1991.

[44] J. Garres-Díaz, E. Fernández-Nieto, and G.Narbona-Reina. A semi-implicit approach for sediment transport models with gravitational effects. *Applied Mathematics and Computation*, 421, 2022.

[45] L. Gascón and J. M. Corderán. Construction of second-order tvd schemes for nonhomogeneous hyperbolic conservation laws. *Journal of Computational Physics*, 172:261–297., 2001.

[46] C. Gauss. Theoria attractionis corporum sphaeroidicorum ellipticorum homogeneorum methodo nova tractata. *Commentationes societatis regiae scientiarium Gottingensis recentiores*, 2:355–378, 1813.

[47] S. Godunov. Finite difference methods for the computation of discontinuous solution of the equation of fluid dynamics. *Mat.Sb*, 47:271–306, 1959.

[48] S. Gottlieb, D. Ketcheson, and C. W. Shu. Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations. *World Scientific, Singapore*, 2011.

[49] S. Gottlieb and C. W. Shu. Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation*, 67(221):73–85, 1998.

[50] A. J. Grass. Sediments transport by waves and currents. *SERC London Cent Mar Technol*, Report No. FL29, 1981.

[51] G. Green. An essay on the application of mathematical analysis to the theories of electricity and magnetism. *Nottingham, England: T. Wheelhouse*, pages 10–12, 1838.

[52] L. Grosheintz-Laval and R. Kappeli. High-order well-balanced finite volume schemes for the Euler equations with gravitation. *Journal of Computational Physics*, 378:324–343., 2019.

[53] L. Grosheintz-Laval and R. Kappeli. Well-balanced finite volume schemes for nearly steady adiabatic flows. *Journal of Computational Physics*, 423:28., 2020.

[54] I. Gómez-Bueno, M. J. Castro-Díaz, C. Parés, and G. Russo. Collocation methods for high-order well-balanced methods for systems of balance laws. *Mathematics*, 9, 2021.

[55] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comp. Phys.*, 49:357–393, 1983.

[56] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy. Uniformaly high order accuracy essential non-oscillatory schemes iii. *J. Comp. Phys.*, 71:231–303, 1987.

[57] J. S. Hesthaven. *Numerical Methods for Conservation laws: From Analysis to Algorithms*. SIAM, 2018.

[58] C. Hirsch. *Numerical computation of internal and external flows (volume 1): fundamentals of numerical discretization*. John Wiley & Sons, Inc., New York, NY, USA, 2007.

[59] C. Hirsch. *Numerical computation of internal and external flows (volume 2): the fundamentals of computational fluid dynamics*. John Wiley & Sons, Inc., New York, NY, USA, 2007.

[60] H. Holden, L. Holden, and R. Hoegh-krohn. A numerical method for first order nonlinear scalar conservation laws in one dimension. *Comput. Math. Appl.*, 15(68):595–602, 1988.

[61] H. Holden and N. H. Risebro. *Front tracking for hyperbolic conservation laws*, volume 152. Applied Mathematical Science, 2 edition, 2015.

[62] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1991.

[63] J. Hudson. Numerical Techniques for Morphodynamic Modelling. *Ph.D. Thesis, Department of Mathematics, The University of Reading, Whiteknigths, Reading,*, 2001.

[64] H. Hugoniot. Sur la propagation du mevement dans les coprs et spécialement dan les gaz parfaits. *J- Ecole Polythecnique*, 53:3–97, 1887.

[65] J. Humpherys and T. J. Jarvis. *Foundations of Applied Mathematics Volume 2: Algorithms, Approximation, Optimization.* Society for Industrial and Applied Mathematics, 2020.

[66] J. Humpherys, T. J. Jarvis, and E. J. Evans. *The Fundamental Theorem of Arithmetic". Foundations of Applied Mathematics Volume 1: Mathematical Analysis*. Society for Industrial and Applied Mathematics, 2017.

[67] G. S. Jiang and C. W. Shu. Efficient implementation of weighted ENO schemes. *J. Comput. Phys.*, 126:202–228, 1996.

[68] F. Kanbar, R. Touma, and C. Klingenberg. Well-balanced central schemes for the one and two-dimensional Euler systems with gravity. *Appl. Numer. Math*, 156:608–626, 2020.

[69] R. Kappeli and S.Mishra. Well-balanced schemes for the Euler equations with gravitation. *Journal of Computational Physics*, 259:199–219, 2014.

[70] F. Kemm. A comparative study of tvd-limiters well-known limiters and an introduction of a new ones. *Int. J. Numer. Methods Fluids*, 67(4):404–440, 2010.

[71] C. Klingberg, G. Puppo, and M. Semplice. Arbitrary order finite volume well-balanced schemes for the Euler equations with gravity. *SIAM J. Sci. Comput.*, 41(2):695–721, 2019.

[72] A. J. Kriel. Error analysis of flux limiter schemes at extrema. *J. Comput. Phys.*, 328:371–386, 2017.

[73] S. N. Kruzkov. First order quasilinear equation with several independent variables. *Math. Sb.*, 81(123):228–255, 1970.

[74] A. Kurganov and E. Tadmor. Solution of two-dimensional Riemann problems for a gas dynamics without Riemann problem solvers. *Numer. Methods Partial Differential Equations*, 18:584–608, 2002.

[75] P. Lax and R. D. Richtmyer. Survey on stability of linear finite difference equations. *Communication on Pure and Applied Mathematics*, 9(2):1956, 267-293.

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

[77] P. Lax and B. Wendroff. Difference schemes for hyperbolic equations with high order accuracy. *Communications Pure and Applied Mathematics*, XVII(2):381–393, 1964.

[78] P. Lax and L. Xu-Dong. Solution of two-dimensional Riemann problems of gas dynamics by positive schemes. *SIAM Journal on Scientific Computing*, 19F(2):319–340, 1998.

[79] P. D. Lax. Hyperbolic systems of conservation laws, II. *CPAM*, 10:537–566, 1957.

[80] P. D. Lax. Shock waves and entropy. *Contribution to nonlinear functional analysis*, Accademic Press:603–634, 1971.

[81] P. LeFloch. *Hyperbolic Systems of Conservation Laws: The Theory of Classical and Nonclassical Shock Waves*. Lectures in Mathematics ETH Zurich Birkhauser-Verlag, 1 edition, 2002.

[82] R. LeVeque. *Numerical Methods for conservation laws*. Springer Basel AG. lectures in Mathematics ETH Zurich., 2 edition, 1992.

[83] R. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press., 1 edition, 2002.

[84] R. LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems (Classics in Applied Mathematics)*. Society for Industrial and Applied Mathematics, Philadelpia, PA. USA., 1 edition, 2007.

[85] D. W. Levy, G. Puppo, and G. Russo. Central weno schemes for hyperbolic systems of conservation laws. *Mathematical Models and Numerical Analysis*, 33:547–571, 1999.

[86] D. W. Levy, G. Puppo, and G. Russo. A fourth order central weno scheme for multidimensional systems of hyperbolic conservation laws. *SIAM J. Scientific Computing*, 24:480–506, 2002.

[87] T. P. Liu. The entropy condition and the admissibility of shocks. *J. Math. Anal. Appl.*, 53:78–88, 1976.

[88] X. D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *Journal of Computational Physics*, 115:200–212, 1994.

[89] R. Loubère, M. Dumbser, and S. Diot. A new family of high order unstructured mood and ader finite volume schemes for multidimensional systems of hyperbolic conservation laws. *Communication in Computational Physics*, 16:718–763, 2014.

[90] M. Lukacova-Medvid'ova and G. Warnecke. Lax-wendroff type second order evolution galerkin methods for multidimensional hyperbolic systems. *East-West J. Numer. Math.*, 8:127–152, 2000.

[91] R. MacCormack. The effect of viscosity in hypervelocity impact cratering. *AIAA Pape, Cincinnati, Ohio*, pages 69–354, 1969.

[92] N. Macon and A. Spitzbart. Inverses of vandermonde matrices. *The American Mathematical Monthly*, 65(2):95–100, 1958.

[93] A. Meister and J. Struckmeier. *Hyperbolic Partial Differential Equation: Theory Numerics and Application*. Friedr Vieweg & Sohn, 1 edition, 2002.

[94] J. Murillo and P. García-Navarro. An exner-based coupled model for two-dimensional transient flow over erodible bed. *Journal of Computational Physics*, 229(23):8704–8732, 2010.

[95] O. Oleinik. Discontinuous solution of nonlinear differential equations. *Amer. Math. Soc. Transl. Ser. 2*, 26:95–172, 1957.

[96] C. Pares. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44:300–321, 2006.

[97] C. Parés and C. Parés-Pulido. Well-balanced high-order finite difference methods for systems of balance laws. *J. Comput. Phys.*, 45:35, 2021.

[98] S. Qian, G. Li, F. Shao, and Q. Niu. Well-balanced central weno schemes for the sediment transport model in shallow water. *Comput. Geosci*, 22(3):763–773, 2018.

[99] J. Qiu and C. W. Shu. Finite difference weno schemes with lax-wendroff-type time discretizations. *SIAM J. Sci. Comput.*, 24(6):2185–2198, 2003.

[100] A. Quarteroni and F. Saleri. *Scientific Computing with MATLAB. Texts in computational science and engineering*. Springer, 2003.

[101] W. J. M. Rankine. On the thermodynamic theory of waves of finite longitudinal disturbance. *Phil. Trans. Roy. Soc. London*, 160:277–288, 1870.

[102] R. D. Richtmyer and K. W. Morton. Difference methods for initial-value problems. *Interscience Tracts in Pure and Appl. Math. Interscience, New York*, 1, 1967.

[103] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*. Krieger, 2 edition, 1994.

[104] K. Riley, M. Hobson, and S. Bence. *Mathematical methods for physics and engineering.* Cambridge University Press, 2010.

[105] P. Roe. Characteristic-based schemes for the Euler equations. *Annu. Rev. Fluid Mech.*, 18:337–365, 1986.

[106] G. Russo and A. Khe. High order well balanced schemes for systems of balance laws. hyperbolic problems: theory, numerics and applications. *Proc. Sympos. Appl. Math.*, 67(2):919–928, 2009.

[107] S. Salsa. *Equazioni a derivate parziali.* Springer, 3 edition, 2016.

[108] T. Schwartzkopff, C. Munz, and E. Toro. Ader: a high-order approach for linear hyperbolic systems in 2d. *J. Sci. Comput.*, 17:231–240, 2002.

[109] C. Shu. Total-variation-diminishing time discretizations. *SIAM Journal of Scientific and Statistical Computing*, 9(6):1073–1084, 1988.

[110] C. W. Shu. Essentially non-oscillatory and weighted essentially non–oscillatory schemes for hyperbolic conservation laws. Technical report, Institute for Computer Applications in Science and Engineering (ICASE), 1997.

[111] C. W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock capturing scheme. *J. Comput. Phys.*, 77:439–471, 1988.

[112] C. W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock capturing scheme. *J. Comput. Phys.*, 81(1):32–78, 1989.

[113] J. Smoller. *Shock Waves and Reaction-Diffusion Equations.* Grundlehren der Mathematischen Wissenschaften. Springer Science & Business Media, 1994.

[114] G. A. Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of Computational Physics*, 27(1):1–31, 1978.

[115] J. C. Strikwerda. *Finite Difference Scheme and Partial Differential Equations.* SIAM, 2 edition, 2004.

[116] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.

[117] A. Thomann, M. Zenk, and C. Klingenberg. A second-order positivity-preserving well-balanced finite volume scheme for euler equations with gravity for arbitrary hydrostatic equilibria. *International Journal for numerical methods in fluids*, 89(11):465–482, 2019.

[118] V. Titarev and E. Toro. Ader: arbitrary high order godunov approach. *J. Sci. Comput.*, 17:609–618, 2002.

[119] E. Toro. Primitive, conservative and adaptive schemes for hyperbolic conservation laws. *Numerical Methods for Wave Propagation. Academic Publishers*, 1:323–385, 1998.

[120] E. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, third edition, 2009.

[121] E. Toro, R. Millington, and L. Nejal. Towards very high order godunov scheme. in e. toro (ed) godunov methods: Theory and applications. *Kluwer/Plerum Academic Publisher*, pages 905–932, 2001.

[122] J. Trangenstein. *Numerical Solution of Partial Differential Equations*. Cambridge Press, 2006.

[123] J. Trangenstein. *Numerical Solution of Hyperbolic Partial Differential Equations*. Cambridge Press, 2009.

[124] S. Wang and Z. Xu. Total variation bounded flux limiters for high order finite difference schemes solving one-dimensional scalar conservation laws. *Math. Comp.*, 88:691–716, 2019.

[125] B. Wendroff. Theoretical numerical analysis. *Academic Press, New York*, 1966.

[126] B. Wendroff. The Riemann problem for materials with nonconvex equation of stare. *J. Math. Anal. Appl.*, 38:454–466, 1972.

[127] J. Witham. Linear and nonlinear waves. *Pure and Applied Mathematics*, 1999.

[128] P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *Journal of Computational Physics*, 1:115–173, 1984.

[129] Y. Xing. Numerical methods for the nonlinear shallow water equations. *Handbook of Numerical Methods, Springer*, 18:361–384., 2017.

[130] Y. Xing and C. W. Shu. High-order well-balanced finite difference WENO schemes with the exact conservation property for the shallow water equations. *Journal of Computational Physics*, 208:206–227., 2006.

[131] Y. Xing and C. W. Shu. High-order well-balanced finite difference WENO schemes for a class of hyperbolic systems with source terms. *Journal of Scientific Computing*, 27:2006, 477-494.

[132] D. Zorío, A. Baeza, and P. Mulet. An approximate lax-wendroff-type procedure for high order accurate scheme for hyperbolic conservation laws. *J. Sci. Comput.*, 71(1):246–273, 2017.

Appendix A

A.1 Numerical Differential Formulas

Let us define two operators D and A such that, given a variable z, we compute the $k-$th numerical derivatives using respectively $(2p+1)$ and $2p-$stencil point as follows:

$$D_P^k(z_*, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-p}^{p} \delta_{p,j}^k z_{i+j}$$
$$A_P^k(z_*, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-p+1}^{p} \gamma_{p,j}^k z_{i+j}.$$

In practise, D is the operator that compute the $k-$th numerical derivative centered at position x_i using $(2p+1)-$point stencil; while, A is the operator that compute the $k-$th numerical derivative centered at position $x_{i+\frac{1}{2}}$ using $2p-$point stencil. Observe that the symbol $*$ indicates with respect to which the operator is applied and, for all $j = -p, \ldots, p$, z_{i+j} is an approximation of $z(x_i + j\Delta x)$. We will also define the $k-$th derivatives in space or time at position q using $2p-$point stencil as:

$$\partial_x^k u(x_i + q\Delta x, t_n) \simeq A_p^{k,q}(u_{i,*}^n, \Delta x) = \frac{1}{\Delta x^k} \sum_{j=-p}^{p} \delta_{p,j}^{k,q} u_{i+j}$$
$$\partial_x^k u(x_i, t_n + q\Delta t) \simeq A_p^{k,q}(u_i^*, \Delta t) = \frac{1}{\Delta x^k} \sum_{r=-p+1}^{p} \gamma_{p,r}^{k,q} u_i^{n+r}.$$

For the sake of simplicity, remembering that δ and γ do not depend on i and Δx, let us suppose $i = 0$, $x_0 = 0$ and $\Delta x = 1$. For this reason, we consider the $k-$th derivative of f

centered in position 0 and q respectively adopting $(2p+1)$ and $2p-$point stencil as

$$f^{(k)}(0) \simeq D_p^k(f,1) = \sum_{j=-p}^{p} \delta_{p,j}^k f(j), \tag{A.1.1}$$

$$f^{(k)}(q) \simeq A_p^{k,q}(f,1) = \sum_{j=-p+1}^{p} \delta_{p,j}^{k,q} f(j). \tag{A.1.2}$$

Observe that eq. (A.1.1) is the interpolatory formulas of $(2p+1)-$points, then it is exact for all polynomials of degree $\leq 2p$. Thus applying operator (A.1.2) to x^s, with $s = 0, \ldots, 2p$ at position $x = 0$, the δ coefficients must satisfy the Vandermonde condition, see Section 6.1 [62] or [43, 92],

$$\sum_{j=-p}^{p} j^k \delta_{p,j}^k = k!, \qquad \sum_{j=-p}^{p} j^s \delta_{p,j}^k = 0, \quad s \neq k \qquad 0 \leq s,k \leq 2p. \tag{A.1.3}$$

In similar way, working with $2p-$points, the γ coefficients must satisfy

$$\sum_{j=-p+1}^{p} j^k \gamma_{p,j}^k = k!, \qquad \sum_{j=-p+1}^{p} j^s \gamma_{p,j}^k = 0, \quad s \neq k \qquad 0 \leq s,k \leq 2p-1. \tag{A.1.4}$$

In practice, we find:

$$\sum_{j=-p+1}^{p} j^k \gamma_{p,j}^k = \begin{cases} 1 & \text{if } k = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{A.1.5}$$

Let us consider a generic function f and $(2p+1)-$points, the Lagrange polynomial [65, 66, 100] $\mathscr{L}(x)$ that interpolate f on $(2p+1)-$points is:

$$\mathscr{L}(x) = \sum_{j=-p}^{p} f(x_{i+j}) \ell_{p,j}(x),$$

where $\ell_{p,j}(x)$ represents the Lagrange basis

$$\ell_{p,j}(x) = \prod_{r=-p, r\neq j}^{p} \frac{x-r}{j-r}, \qquad -p \leq j \leq p. \tag{A.1.6}$$

The $k-$th derivatives of lagrangian basis $\ell_{p,j}$ represent the $\gamma_{p,j}^k$ coefficients. In fact, it is enough write $\ell_{p,j}$ as:

$$\ell_{p,j} = \prod_{r=-p, r\neq j}^{p} \frac{x-r}{j-r} = \sum_{k=0}^{2p} \frac{\delta_{p,j}^k}{k!} x^k,$$

where the last equality is guaranteed by Taylor expansion at $x = 0$.

Proposition A.1.1 *The coefficients $\delta^k_{p,j}$ introduced on* (A.1.1) *satisfy:*

$$\delta^k_{p,j} = (-1)^k \delta^k_{p,-j}; \tag{A.1.7}$$

$$\delta^k_p = 0 \quad \text{if } k \text{ is odd}; \tag{A.1.8}$$

$$\sum_{j=-p}^{p} \delta^k_{p,j} j^{(2p+1)} = 0 \quad \text{if } k \text{ is even}; \tag{A.1.9}$$

$$\sum_{j=-p}^{p} \delta^k_{p,j} j^{(2p+2)} = 0 \quad \text{if } k \text{ is odd}. \tag{A.1.10}$$

Proof. (A.1.7) comes out from:

$$\ell_{p,-j}(x) = \ell_{p,j}(-x).$$

(A.1.8) follows from (A.1.7). (A.1.9) can be written as

$$\sum_{j=-p}^{p} \delta^k_{p,j} j^{(2p+1)} = \sum_{j=-p}^{-1} \delta^k_{p,j} j^{(2p+1)} + \sum_{j=1}^{p} \delta^k_{p,j} j^{(2p+1)} = 0$$

since k is even and $2p+1$ is odd. In similar way (A.1.10). ■

An important property to written the numerical scheme in conservative form is the next proposition.

Proposition A.1.2 *For $k \geq 1$ the following relations are satisfied:*

$$\delta^k_{p,p} = \gamma^{k-1,\frac{1}{2}}_{p,p}; \tag{A.1.11}$$

$$\delta^k_{p,j} = \gamma^{k-1,\frac{1}{2}}_{p,j} - \gamma^{k-1,\frac{1}{2}}_{p,j+1} \quad j = -p+1, \ldots, p-1; \tag{A.1.12}$$

$$\delta^k_{p,-p} = -\gamma^{k-1,\frac{1}{2}}_{p,-p+1}. \tag{A.1.13}$$

Proof. First of all, let us observe that given a polynomial g of degree 1,

$$g'(0) = g\Big(\frac{1}{2}\Big) - g\Big(-\frac{1}{2}\Big).$$

Now, considering the following formulas:

$$f^{(k-1)}\Big(\frac{1}{2}\Big) \simeq A_p^{k-1,\frac{1}{2}}\Big(f,1\Big) = \sum_{j=-p+1}^{p} \gamma_{p,j}^{k-1,\frac{1}{2}} f(j); \tag{A.1.14}$$

$$f^{(k-1)}\Big(-\frac{1}{2}\Big) \simeq A_p^{k-1,\frac{1}{2}}\Big(f_{-1},1\Big) = \sum_{j=-p+1}^{p} \gamma_{p,j}^{k-1,\frac{1}{2}} f(j-1). \tag{A.1.15}$$

These formulas are the interpolatory formulas adopting $2p$ nodes hence the degree are $2p-1$. Then, formulas (A.1.14) and (A.1.15) are exactly when applied to polynomial with degree less or equal then $2p-1$. Let us consider f a polynomial of degree $2p$,

$$f^{(k)}(0) = f^{(k-1)}\Big(\frac{1}{2}\Big) - f^{(k-1)}\Big(-\frac{1}{2}\Big) = A_p^{k-1,\frac{1}{2}}\Big(f,1\Big) - A_p^{k-1,\frac{1}{2}}\Big(f_{-1},1\Big). \tag{A.1.16}$$

Then,

$$\begin{aligned} f^{(k)}(0) = \gamma_{p,p}^{k-1,\frac{1}{2}} f(p) + \Big(\gamma_{p,p-1}^{k-1,\frac{1}{2}} - \gamma_{p,p}^{k-1,\frac{1}{2}}\Big) f(p-1) + \dots \\ + \Big(\gamma_{p,-p+1}^{k-1,\frac{1}{2}} - \gamma_{p,-p+2}^{k-1,\frac{1}{2}}\Big) f(-p+1) - \gamma_{p,-p+1}^{k-1,\frac{1}{2}} f(-p). \blacksquare \end{aligned}$$

Proposition A.1.3 *Given $1 \le k \le 2p-1$ and $0 \le s \le k$ we get that:*

$$\sum_{j=-p+1}^{p} \gamma_{p,j}^{s,q} \gamma_{p,l}^{k-s,j} = \gamma_{p,l}^{k,q}, \quad l = -p+1, \dots, p. \tag{A.1.17}$$

Proof. The proof is similar to the one of the preceding Proposition A.1.2. Indeed, let us consider the following formula

$$f^{(k)} \simeq \sum_{j=-p+1}^{p} \gamma_{p,j}^{s,q} f_j^{(k-s)},$$

where

$$f_j^{(k-s)} = \sum_{l=-p+1}^{p} \gamma_{p,l}^{k-s,j} f(l).$$

Hence,

$$f^{(k)}(q) \simeq \sum_{l=-p+1}^{p} \left(\sum_{j=-p+1}^{p} \gamma_{p,j}^{s,q} \gamma_{p,l}^{k-s,j} \right) f(l). \blacksquare$$

Appendix B

B.1 Compact Approximate Taylor properties

Lemma B.1.1 *Let be $k > 1$ and $f_{i,j}^{(k-1)}$ defined as in Section 3.1.3. Then $f_{i,j}^{(k-1)} = aU_{i,j}^{(k-1)}$ when $f(U) = aU$.*

Proof.

$$
\begin{aligned}
f_{i,j}^{(k-1)} &= \frac{1}{\Delta t^{k-1}} \sum_{r=-P+1}^{P} \gamma_{P,r}^{k-1,0} f_{i,j}^{k-1,n+r} = \\
&= \frac{a}{\Delta t^{k-1}} \sum_{r=-P+1}^{P} \gamma_{P,r}^{k-1,0} \Big(U_{i+j}^{n} + \sum_{\ell=1}^{k-1} \frac{(r\Delta t)^{\ell}}{\ell!} U_{i,j}^{(\ell)} \Big) = \\
&= \frac{a}{\Delta t^{k-1}} \left(\Big(\sum_{r=-P+1}^{P} \gamma_{P,r}^{k-1,0} \Big) U_{i+j}^{n} + \sum_{\ell=1}^{k-1} \frac{\Delta t^{\ell}}{\ell!} \Big(\sum_{r=-P+1}^{P} \gamma_{P,r}^{k-1,0} r^{\ell} \Big) U_{i,j}^{(\ell)} \right) \\
&= aU_{i,j}^{(k-1)},
\end{aligned}
$$

where the last identity is satisfied from (A.1.4). ■

Theorem B.1.1 *The Compact Approximate Taylor method is a properly generalization of the high order Lax-Wendroff scheme* (2.2.1) *for linear systems of conservation laws.*

Proof. In order to prove that CAT2P is a properly generalization of the $2P-$order Lax-Wendroff method we have to prove that CAT2P reduces to (2.2.1) when applied to systems

(2.1.2). For this reason, given $k > 1$, $f_{i,j}^{(k-1)} = aU_{i,j}^{(k-1)}$ (Lemma B.1.1). On the other side,

$$\begin{aligned}
U_{i,j}^{(k)} &= -\frac{1}{\Delta x}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}U_{i,s}^{(k-1)} = \\
&= -\frac{a}{\Delta x}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}U_{i,s}^{(k-1)} = \\
&= \frac{a^2}{\Delta x^2}\sum_{s=-P+1}^{P}\gamma_{P,s}^{1,j}\sum_{\ell=-P+1}^{P}\gamma_{P,\ell}^{1,s}U_{i,\ell}^{(k-2)} \\
&= \frac{a^2}{\Delta x^2}\sum_{s=-P+1}^{P}\gamma_{P,s}^{2,j}U_{i,s}^{(k-2)} = \ldots = \\
&= \frac{(-1)^k a^k}{\Delta x^k}\sum_{s=-P+1}^{P}\gamma_{P,s}^{k,j}U_{i+j}^{n},
\end{aligned}$$

where Proposition A.1.3 has been used. In addition,

$$\begin{aligned}
A_P^{0,\frac{1}{2}}\Big(f_{i,*}^{(k-1)},\Delta x\Big) &= \frac{1}{\Delta x}\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,\frac{1}{2}}f_{i,j}^{(k-1)} = \\
&= -\frac{a}{\Delta x}\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,\frac{1}{2}}U_{i,j}^{(k-1)} = \\
&= \frac{(-1)^{k-1}a^k}{\Delta x^k}\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,\frac{1}{2}}\sum_{s=-P+1}^{P}\gamma_{P,s}^{k-1,j}U_{i+s}^{n} = \\
&= \frac{(-1)^{k-1}a^k}{\Delta x^k}\sum_{s=-P+1}^{P}\Big(\sum_{j=-P+1}^{P}\gamma_{P,j}^{0,\frac{1}{2}}\gamma_{P,s}^{k-1,j}\Big)U_{i+s}^{n} = \\
&= \frac{(-1)^{k-1}a^k}{\Delta x^k}\sum_{s=-P+1}^{P}\gamma_{P,s}^{k-1,\frac{1}{2}}U_{i+j}^{n} = \\
&= \frac{(-1)^{k-1}a^k}{\Delta x^k}A_P^{k-1,\frac{1}{2}}U_{*}^{n},
\end{aligned}$$

where Proposition A.1.3 has been used. In this way, the flux reconstruction of order $2P$ (3.1.2) becomes:

$$\begin{aligned}
F_{i+\frac{1}{2}}^{P} &= \sum_{k=1}^{2P}\frac{\Delta t^{k-1}}{k!}A_P^{0,\frac{1}{2}}\Big(f_{i,*}^{(k-1)},\Delta x\Big) = \\
&= \sum_{k=1}^{2P}(-1)^{k-1}\frac{a^k\Delta t^{k-1}}{k!}A_P^{k-1,\frac{1}{2}}\Big(U_{*}^{n},\Delta x\Big).
\end{aligned}$$

■

Remark B.1.1 *As consequence of Theorem B.1.1, the Compact Approximate Taylor Method (CAT2P) properly extends the $2P-$ order Lax-Wendroff scheme when applied to linear systems of conservation laws. Thus, the CAT method is linearly stable (in the L^2 sense) under the usual CFL-condition (see also [82]-[84])*

$$\max_i(|f'(U_i)|)\frac{\Delta t}{\Delta x} \leq 1.$$

Theorem B.1.2 *The Compact Approximate Taylor scheme is a $2P-$order method.*

Proof. In order to prove that CAT has order $2P$, let us consider an exact solution $U(x,t)$ sufficiently smooth. The idea is to perform a step of the method starting from a generic point value at time t_n, $U(x_i,t_n)$, and prove that the difference of the approximation of the exact solution with two consecutive time step has order $2P$. For this reason, let us consider an approximation of the first time derivative $U^{(1)}_{i,j}$ for all $j=-P+1,\dots,P$,

$$\begin{aligned} U^{(1)}_{i,j} &= -A^{1,j}_P\left(f^{(0)}_{i,*},\Delta x\right) = -\partial_x f(U)(x_{i+j},t_n) + O(\Delta x^{2P-1}) = \\ &= \partial_x f(U)(x_{i+j},t_n) + O(\Delta x^{2P-1}). \end{aligned}$$

Let be $P^1_{i,j}(s)$ the Taylor expansion polynomial truncated at first term $P^1_{i,j}(s) = U(x_{i+j},t_n) + sU^{(1)}_{i,j}$. For construction,

$$f^{1,n+r}_{i,j} = f\left(U(x_{i+j},t_n) + r\Delta t U^{(1)}_{i,j}\right) = f\left(P^1_{i,j}(r\Delta t)\right) + O(\Delta x^{2P}).$$

The approximation of the first time derivative of flux has order $O(\Delta x^{2P-1})$. Indeed,

$$
\begin{aligned}
f_{i,j}^{(1)} &= A_P^{1,0}\Big(f_{i,j}^{1,*}, \Delta t\Big) = \\
&= \frac{1}{\Delta t}\sum_{r=-P+1}^{P} \gamma_{P,j}^{1,0} f_{i,j}^{1,n+r} = \\
&= \frac{1}{\Delta t}\sum_{r=-P+1}^{P} \gamma_{P,j}^{1,0} f\Big(P_{i,j}^{1}(r\Delta t)\Big) + O(\Delta x^{2P}) = \\
&= \frac{1}{\Delta t}\sum_{r=-P+1}^{P} \gamma_{P,j}^{1,0} \sum_{k=0}^{2P-1} \frac{1}{k!} d^k(f \circ P_{i,j}^{1})(t_n) r^k \Delta t^k + O(\Delta x^{2P}) = \\
&= \frac{1}{\Delta t}\sum_{k=0}^{2P-1} \frac{1}{k!} d^k(f \circ P_{i,j}^{1})(t_n) \Delta t^k \sum_{r=-P+1}^{P} \gamma_{P,j}^{1,0} r^k + O(\Delta x^{2P}) = \\
&= d^1(f \circ P_{i,j}^{1})(t_n) + O(\Delta x^{2P}) = \\
&= \partial_t f(U)(x_{i+j}, t_n) + O(\Delta x^{2P}),
\end{aligned}
$$

where (A.1.4) has been used. The previous idea should be extend to every k between 1 and $2P-1$ in the following way:

$$
f_{i,j}^{(k)} = \partial_t^k f(U)(x_{i+j}, t_n) + O(\Delta x^{2P-1}).
$$

Linking all the results we obtain:

$$
\begin{aligned}
&U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) + \frac{\Delta t}{\Delta x}\Big(F^P_{i+\frac{1}{2}} - F^P_{i-\frac{1}{2}}\Big) = \\
&\quad = U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) + \frac{1}{\Delta x}\sum_{k=1}^{2P}\frac{\Delta t^k}{k!}\Big(A_P^{0,\frac{1}{2}}\big(f_{i,*}^{(k-1)},\Delta x\big) - A_P^{0,\frac{1}{2}}\big(f_{i-1,*}^{(k-1)},\Delta x\big) = \\
&\quad = U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) \\
&\qquad + \frac{1}{\Delta x}\sum_{k=1}^{2P}\frac{\Delta t^k}{k!}\Big(A_P^{0,\frac{1}{2}}\big(\partial_t^{k-1}f(U),\Delta x\big) - A_P^{0,\frac{1}{2}}\big(\partial_t^{k-1}f(U),\Delta x\big) + O(\Delta x^{2P+1}) = \\
&\quad = U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) + \frac{1}{\Delta x}\sum_{k=1}^{2P}\frac{\Delta t^k}{k!}D_P^1\big(\partial_t^{k-1}f(U),\Delta x\big) + O(\Delta x^{2P+1}) = \\
&\quad = U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) + \frac{1}{\Delta x}\sum_{k=1}^{2P}\frac{\Delta t^k}{k!}\partial_t^{k-1}f(U)(x_i,t_n) + O(\Delta x^{2P+1}) = \\
&\quad = U(x_{i+j},t_{n+1}) - U(x_{i+j},t_n) - \frac{1}{\Delta x}\sum_{k=1}^{2P}\frac{\Delta t^k}{k!}\partial_t^{k}U(x_i,t_n) + O(\Delta x^{2P+1}) = \\
&\quad = O(\Delta x^{2P+1}).
\end{aligned}
$$

■

Appendix C

C.1 Numerical Coefficients

The coefficients $\delta_{P,j}^{k,q}$ and $\gamma_{P,j}^{k,q}$ of the differentiation formulas (A.1.2) and (A.1.11)-(A.1.13) for $P = 1, 2, 3$ are shown in Figures C.1.1 and C.1.2 respectively. Algorithms to compute those coefficients can be found in [41] and [15].

	q	k	j = -2	j = -1	j = 0	j = 1	j = 2	j = 3
p = 1	1/2	0			1/2	1/2		
		1			-1	1		
p = 2	1/2	0		-0	4/7	4/7	-0	
		1		0	- 5/4	1 1/4	-0	
		2		1/2	- 1/2	- 1/2	1/2	
		3		-1	3	-3	1	
p = 3	1/2		0	- 1/7	5/8	5/8	- 1/8	0
			-0	1/7	-1 1/3	1 1/3	- 1/7	0
			- 1/8	7/8	- 3/4	- 3/4	7/8	- 1/8
			1/6	-1 5/6	4 2/3	-4 2/3	1 5/6	- 1/6
			1/2	-1 1/2	1	1	-1 1/2	1/2
			-1	5	-10	10	-5	1

Figure C.1.1: The $\delta_{P,j}^{k,q}$ coefficients of the differentiation formula (A.1.2) for $P = 1, 2, 3$.

P = 1

q	k	j = 0	j = 1
0	0	1	0
	1	-1	1
1	0	0	1
	1	-1	1

P = 2

q	k	j = -1	j = 0	j = 1	j = 2
-1	0	1	0	0	0
	1	- 11/6	3	- 3/2	1/3
	2	2	-5	4	-1
	3	-1	3	-3	1
0	0	0	1	0	0
	1	- 1/3	- 1/2	1	- 1/6
	2	1	-2	1	0
	3	-1	3	-3	1
1	0	0	0	1	0
	1	1/6	- 1/1	1/2	1/3
	2	0	1	-2	1
	3	-1	3	-3	1
2	0	0	0	0	1
	1	- 1/3	3/2	- 3/1	11/6
	2	-1	4	-5	2
	3	-1	3	-3	1

P = 3

q	k	j = -2	j = -1	j = 0	j = 1	j = 2	j = 3
-2	0	1	0	0	0	0	0
	1	- 137/60	5/1	- 5/1	10/3	- 5/4	1/5
	2	15/4	- 77/6	107/6	- 13/1	61/12	- 5/6
	3	- 17/4	71/4	- 59/2	49/2	- 41/4	7/4
	4	3	-14	26	-24	11	-2
	5	-1	5	-10	10	-5	1
-1	0	0	1	0	0	0	0
	1	- 1/5	- 13/12	2/1	- 1/1	1/3	- 1/20
	2	5/6	- 5/4	- 1/3	7/6	- 1/2	1/12
	3	- 7/4	25/4	- 17/2	11/2	- 7/4	1/4
	4	2	-9	16	-14	6	-1
	5	-1	5	-10	10	-5	1
0	0	0	0	1	0	0	0
	1	1/20	- 1/2	- 1/3	1/1	- 1/4	1/30
	2	- 1/12	4/3	- 5/2	4/3	- 1/12	0
	3	- 1/4	- 1/4	5/2	- 7/2	7/4	- 1/4
	4	1	-4	6	-4	1	0
	5	-1	5	-10	10	-5	1
1	0	0	0	0	1	0	0
	1	- 1/30	1/4	-1	1/3	1/2	- 1/20
	2	0	- 1/12	4/3	- 5/2	4/3	- 1/12
	3	1/4	- 7/4	7/2	- 5/2	1/4	1/4
	4	0	1	-4	6	-4	1
	5	-1	5	-10	10	-5	1
2	0	0	0	0	0	1	0
	1	1/20	- 1/3	1	- 2/1	13/12	1/5
	2	1/12	- 1/2	7/6	- 1/3	- 5/4	5/6
	3	- 1/4	7/4	- 11/2	17/2	- 25/4	7/4
	4	-1	6	-14	16	-9	2
	5	-1	5	-10	10	-5	1
3	0	0	0	0	0	0	1
	1	- 1/5	5/4	- 10/3	5/1	- 5/1	137/60
	2	- 5/6	61/12	- 13/1	107/6	- 77/6	15/4
	3	- 7/4	41/4	- 49/2	59/2	- 71/4	17/4
	4	-2	11	-24	26	-14	3
	5	-1	5	-10	10	-5	1

Figure C.1.2: The $\gamma_{P,j}^{k,q}$ coefficients of the differentiation formulas (A.1.11)-(A.1.13) for $P = 1, 2, 3$.

Acknowledgements

At the end of my doctoral course, I would like to thank my *Advisor*, Prof. Giovanni Russo, that with his valuable suggestions and his constant attention has guided me during these 3 years giving me the opportunity to know new academic and human realities. It is not trivial to find a good Tutor who nurtures a real passion for mathematics and its applications.
A special thanks goes to Prof. Carlos Parés who, with his constant work and his assiduous presence, has marked in a pleasant way my studies and my life.
Thanks to Hugo Carrillo and the whole Malaga group Manolo, Tomás, María Luz, Cipriano, Ernesto G., Ernesto P., Irene, Juan Carlos and Kleiton for making me feel at home.
I would like to thank the *Head of the doctoral school*, Prof.ssa Maria Carmela Lombardo for the efficiency, perseverance, punctuality and commitment with which she has followed us in these years.
A special thanks goes to my friend Umberto Guarnotta who, through his problems ("problemucci") based on everyday life, filled many boards, sheets and days. Thanks also to my colleagues: Giovanni Nastasi, Giuseppe Pipitò and Clarissa Astuto.
I would also thanks Professors Sebastiano Boscarino, Salvatore Angelo Marano and Giuseppe Di Fazio for the suggestions given to me in these years.
A special thanks goes to my friends Vincenzo, Melania, Saro, Riccardo, Gabriele, Loriana, Alessandra, Giacomo, Francesco, Carolina and Andrea with whom I spent unforgettable moments.
I want to thank my parents and my family members for always supporting me in life and academic choices.
Last but not least, I want to thank in a special way my life and adventures partner Martina who, most of all, was able to understand and support me in difficult times. Thanks to Martina, I had the courage to experiment with new ideas, to put myself in the game and to understand that, after all, obstacles exist to be overcome.