



**HAL**  
open science

# Contributions to the development of residual discretizations for hyperbolic conservation laws with application to shallow water flows

Mario Ricchiuto

► **To cite this version:**

Mario Ricchiuto. Contributions to the development of residual discretizations for hyperbolic conservation laws with application to shallow water flows. Numerical Analysis [math.NA]. Université Sciences et Technologies - Bordeaux I, 2011. tel-00651688v1

**HAL Id: tel-00651688**

**<https://theses.hal.science/tel-00651688v1>**

Submitted on 15 Dec 2011 (v1), last revised 7 Mar 2016 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

---

**Contributions to the development  
of residual discretizations  
for hyperbolic conservation laws  
with application to shallow water flows**

---

---

Manuscript submitted in fulfillment of  
the requirements for the obtention of the  
Habilitation à Diriger des Recherches (HDR)

by

**Mario RICCHIUTO**

---

HDR thesis Committee :

<b>Prof. R. Abgrall</b>	(Research Director)
<b>Prof. M. Azaiez</b>	(Invited Member)
<b>Prof. T. Colin</b>	(Reviewer)
<b>Prof. H. Deconinck</b>	(Invited Member)
<b>Prof. A. Ern</b>	(Reviewer)
<b>Prof. A. Lerat</b>	(President)
<b>Prof. L. Mieussens</b>	(Invited Member)
<b>Prof. C-D. Munz</b>	(Reviewer)

---



# Acknowledgements

I wish to express my gratitude to the reviewers Prof. Munz, Prof. Ern, and Prof. Colin. A warm thank you also to Prof. Abgrall, Prof. Azaiez, Prof. Deconick, Prof. Lerat and Prof. Mieussens, for accepting my invitation to be members of the jury.

Thanks to those I have had the honor to work with, or simply share a discussion in front of a blackboard (too many to name them all), and of course to all my friends and those who withstand me day by day.

Of course, I would not be here writing this if it wasn't for my parents and sister. Sono fortunatissimo ad avere voi che sapete sempre ricordarmi che l'amore non conosce distanza ... Grazie...





# Contents

<b>1</b>	<b>Overview</b>	<b>9</b>
1.1	Residual schemes for hyperbolic conservation laws . . . . .	10
1.1.1	Problem setting . . . . .	10
1.1.2	A (very) short introduction to Residual Distribution . . . . .	10
1.1.3	Limitations . . . . .	14
1.1.4	The appeal of the residual approach . . . . .	17
1.1.5	Relations with other techniques . . . . .	18
1.1.6	My contributions . . . . .	19
1.2	High order schemes for free surface flows . . . . .	21
1.2.1	Problem setting . . . . .	21
1.2.2	Well balancing . . . . .	22
1.2.3	Positivity preservation and wetting/drying . . . . .	23
1.2.4	My contributions . . . . .	23
<b>I</b>	<b>Residual discretizations for hyperbolic conservation laws</b>	<b>27</b>
<b>2</b>	<b>Introduction to RD</b>	<b>31</b>
2.1	Generalities and notation . . . . .	31
2.2	Fluctuation splitting on linear triangles . . . . .	33
2.2.1	Boundary conditions . . . . .	34
2.2.2	Positivity and discrete maximum principle . . . . .	35
2.2.3	Linearity preservation and accuracy . . . . .	35
2.2.4	Multidimensional upwinding . . . . .	38
2.2.5	Scheme zoology : linear schemes . . . . .	41
2.2.6	Scheme zoology : nonlinear schemes and limiters . . . . .	44
2.3	Nonlinear conservation laws and systems . . . . .	48
2.3.1	Conservative linearizations . . . . .	48
2.3.2	Wave decompositions and matrix distribution . . . . .	50
2.4	Time dependent problems . . . . .	51
2.5	Beyond second order of accuracy . . . . .	52
<b>3</b>	<b>High order schemes for steady problems</b>	<b>55</b>
3.1	Conservation via direct flux approximation . . . . .	55
3.1.1	Boundary integration of the flux and flux approximation . . . . .	57
3.1.2	First order and high order schemes . . . . .	58

3.1.3	Additional observations . . . . .	60
3.2	Higher order schemes . . . . .	63
3.2.1	Generalities . . . . .	63
3.2.2	Multidimensional upwind schemes . . . . .	65
3.2.3	Non-upwind higher order schemes . . . . .	72
<b>4</b>	<b>High order schemes for unsteady problems</b>	<b>83</b>
4.1	Accuracy and time dependent conservation laws . . . . .	83
4.1.1	RD prototype for time dependent solutions . . . . .	83
4.1.2	Accuracy analysis . . . . .	84
4.1.3	Scheme zoology . . . . .	88
4.2	High order space-time formulations . . . . .	89
4.2.1	Space-time schemes on triangles and tets . . . . .	89
4.2.2	Space-time schemes on extruded prisms . . . . .	94
4.2.3	Space-time schemes with discontinuous representation in time . . . . .	100
4.3	Schemes based on implicit time-stepping . . . . .	103
4.3.1	Nonlinear schemes : survey and comparison . . . . .	106
4.4	Genuinely explicit schemes . . . . .	111
4.4.1	Digression : on RD and mass matrices . . . . .	112
4.4.2	Step 1 : stabilized Galerkin and explicit RK integration . . . . .	116
4.4.3	Step 2 : inaccurate residuals and stabilization . . . . .	117
4.4.4	Step 3 : mass lumping . . . . .	120
4.4.5	Fluctuations and signals .... . . . .	120
4.4.6	Results . . . . .	121
<b>II</b>	<b>Well-balanced discretizations for shallow water flows</b>	<b>123</b>
<b>5</b>	<b>Challenges in shallow water simulation</b>	<b>127</b>
5.1	Numerical challenges : well balancedness/C-property . . . . .	128
5.2	Numerical challenges : wetting/drying . . . . .	130
<b>6</b>	<b>C-properties</b>	<b>131</b>
6.1	Preliminaries . . . . .	131
6.1.1	Super consistency analysis . . . . .	131
6.2	Basic C-property . . . . .	138
6.3	Generalizations of the C-property . . . . .	140
6.3.1	Constant total energy flows . . . . .	140
6.3.2	Flows in sloping channels with friction . . . . .	143
6.3.3	Flows in sloping channels with transverse bed variations . . . . .	145
<b>7</b>	<b>Wetting/drying</b>	<b>147</b>
7.1	Positivity preservation conditions . . . . .	147
7.2	Wetting/drying and bathymetry approximation . . . . .	148
7.3	Results . . . . .	149
7.3.1	Thacker's oscillations . . . . .	150
7.3.2	Wave run up on a conical island . . . . .	151
7.3.3	The 1993 Okushiri tsunami test case . . . . .	152

<b>III</b>	<b>Conclusions and perspectives</b>	<b>155</b>
<b>8</b>	<b>Summary of contributions and perspectives</b>	<b>157</b>
8.1	Summary of contributions . . . . .	157
8.2	Perspectives : residual schemes for conservation laws . . . . .	158
8.2.1	Schemes for unsteady conservation laws . . . . .	161
8.3	Perspectives : free surface flows and other activities . . . . .	163
8.3.1	High order numerical modeling of free surface flows . . . . .	163
8.3.2	Numerical modeling of oxidation and healing processes in composite materials	165
	<b>Publications by the author</b>	<b>167</b>
	<b>Bibliography</b>	<b>173</b>



# Chapter 1

## Introduction : context and overview of main contributions

This manuscript summarizes a dozen years of research on numerics for hyperbolic conservation laws and application to compressible gas dynamics and Shallow Water flows. This activity started with my arrival as a young undergraduate at the von Karman Institute for Fluid Dynamics in 1999, where I have been introduced to the Rubik's cube problem of constructing a residual based scheme for time dependent hyperbolic conservation laws, possibly upwind, possibly high order, possibly explicit, and possibly yielding oscillation free solutions.

After almost twelve years one of the core topics of my research is : the construction of a residual based scheme for time dependent hyperbolic conservation laws, possibly upwind, possibly high order, possibly explicit, and possibly yielding oscillation free solutions... but I do feel a little closer to the objective than I did in 1999.

In this introductory chapter I will try to put my work into context, and describe and justify the reason behind certain choices and recall what my contributions have been, what is their impact, and on what community. The style of the chapter is willingly informal, not only in the tone, but in the sense that most of the maths are presented in parts I and II, while here I try to discuss ideas and put things in a historical/scientific context. In other words, this chapter has to give the overall perspective and motivation for my work, as well as an overview of the main scientific contributions. It has to answer the questions of why I have been looking into certain subjects, what I have been looking for, and what answers I have been able to come up with.

The chapter is organized in two sections, one entirely devoted to my work on schemes (section §1.1), the other discussing the work on shallow water simulations (section §1.2)

The two remaining parts of the manuscript will be devoted to a more detailed discussion of my contributions. In particular :

- part I discusses my work on the research of novel discretization techniques for hyperbolic conservation laws. The idea is to draw some of the *fundamental principles* underpinning my research, and give a more detailed description for a few selected results ;

- part II discusses the *application of some of these fundamental principles* to the simulation of shallow water flows. Shallow Water simulations represent a road to real life applications on one hand, and on the other allow to push the limits of those fundamental principles that guide my developments in numerics. Details on the main results and computational examples are discussed ;

For a better understanding of the results discussed, both part I and part II begin with an introductory chapter containing as much background information as possible. A short conclusive part ends the manuscript.

Most importantly, the work summarized in this manuscript is not only my own but it is the result of the collaborations with all my excellent colleagues without whom I would be lost. These collaborations are recalled and given credit.

## 1.1 Residual schemes for hyperbolic conservation laws

### 1.1.1 Problem setting

Let us consider the numerical approximation of solutions to the system of nonlinear Partial Differential Equations (PDEs) :

$$\partial_t u + \nabla \cdot \mathcal{F}(u) = 0 \quad \text{on} \quad \Omega \times [0, T_{\text{fin}}] \subset \mathbb{R}^d \times \mathbb{R}^+ \quad (1.1)$$

The *conservation law* (1.1) is assumed hyperbolic and in particular it enjoys all the classical properties (see e.g. [231] and [21, 246, 245, 125, 184, 134]) such as existence of an entropy pair, symmetrizability, and the projection of flux Jacobian (which in general is a  $d$ -dimensional vector of matrices) on the arbitrary  $\vec{\xi} \in \mathbb{R}^d$  direction

$$a_\xi = \vec{a}(u) \cdot \vec{\xi} = \partial_u \mathcal{F}(u) \cdot \xi \quad (1.2)$$

is diagonalizable with linear eigenvalues and linearly independent eigenvectors. Moreover, being nonlinear, (1.1) admits discontinuous solutions compatible with the entropy conditions, and with the Rankine-Hugoniot relations [231].

Suppose that we have an unstructured grid of the spatial domain  $\Omega$ . Let  $\Omega_h$  denote the grid,  $K$  denote the generic grid cell, and let small italic letters (e.g.  $i, j, k$  etc.) denote the nodes of a cell. We seek approximations of  $u$ , the solution of (1.1), on  $\Omega_h$ .

### 1.1.2 A (very) short introduction to Residual Distribution

There exist quite a number of numerical techniques to deal with (1.1). The family of discretizations known as *Fluctuation Splitting* (FS) or *Residual Distribution* (RD) schemes is part of one of the attempts started in the 90's to overcome the accuracy limitations of Finite Volume (FV) schemes<sup>1</sup> based on the solution of pseudo-one dimensional Riemann Problems, while retaining the *monotonicity preservation* philosophy, in contrast to the Finite Element (FE) schemes existing at that time.

---

<sup>1</sup>...and still very much not ended if one is to judge from the number of journal publications and funded research contracts on the topic, but especially on the fact that still most of the codes routinely used in industry are FV-based

The initial idea is due to Phil Roe and is related to his interpretation of flux difference splitting in the context of *fluctuations and signals* [215, 218, 219] : the flux difference being a fluctuation, and its split components being the signals generating variations in time of the discrete solution in cell nodes. In this interpretation, nodal values of the flux are uniquely defined. So, *not only the point of view is shifted from evolving cell averages to evolving nodal values, but the approximation becomes continuous across cell boundaries*. This simple idea spawned a whole family of multidimensional schemes based on the strategy *compute flux balance-split to cell nodes-evolve nodal values* (cf. figure 1.1). In this approach, a continuous interpolation (typically standard  $P^1$  Lagrange triangles) is employed, so that the cell flux balance is well defined (as the flux nodal values in the 1d case).

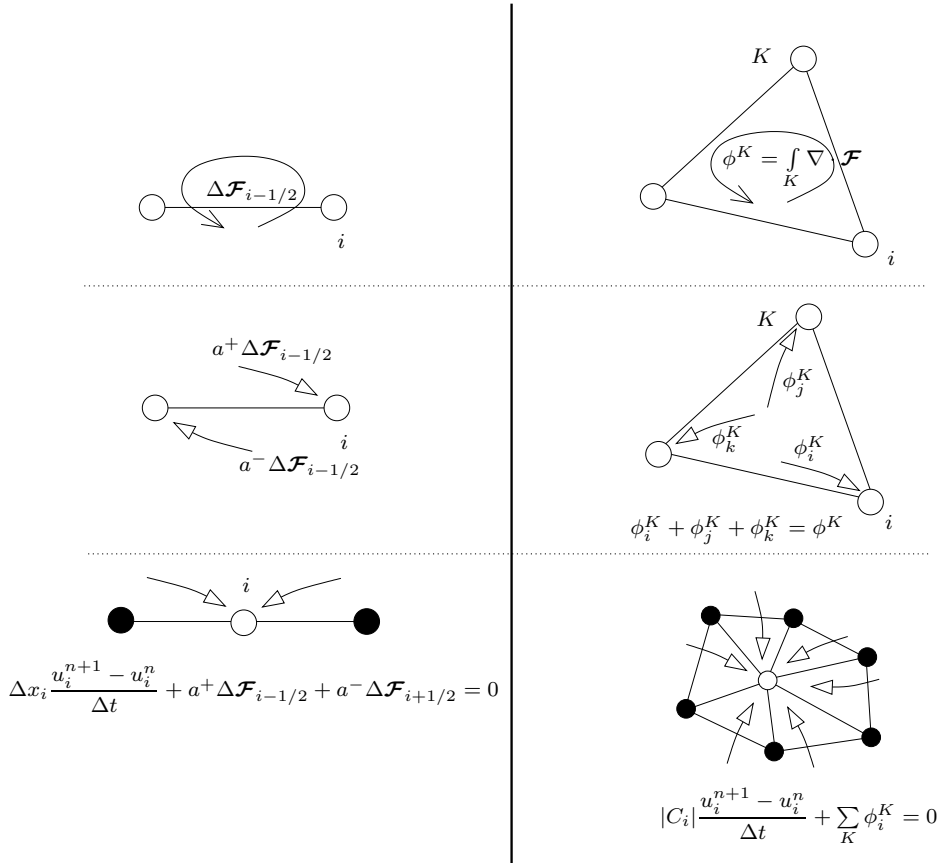


Figure 1.1: The *compute balance-split-evolve* strategy

In the beginning, this new discretization approach was pushed forward mainly by two groups : one at the University of Michigan, led by P.L.Roe ; one at the von Karman Institute for Fluid Dynamics, led by H.Deconinck. The initial developments attracted several other groups that joined later on<sup>2</sup>, including : the University of Reading, and then the University of Leeds (M. Hubbard and co-workers) in UK ; the Politecnico di Bari and later the Università della Basilicata (M. Napolitano and collaborators) in Italy ; in US, apart from the work of

<sup>2</sup>and/or for some time..



Phil Roe and Bram van Leer, some work at NASA (T.J. Barth, W.A. Wood and W.L. Kleb) and ICASE (D.Sidilkover, later at Soreq NRC, Propulsion Physics Laboratory in Israel), and much more recently at Brown University (C.-W. Shu), at the NIA (H.Nishikawa), and at the University of Wisconsin (J. A.Rossmanith) ; in Canada the UTIAS University of Toronto (C.P.T.Growth and S.Guzik) ; some important contributions from Lund University in Sweden (D.A. Caraeni) ; in France some early work by Perthame, some developments by DASSAULT, and more recently a lot of input coming from INRIA (R.Abgrall and collaborators) ; the Technical University of Lisbon in Portugal (L.M.C.Gato and co-workers) ; the Yeditepe University in Turkey (N.Aslan) ; and others.

In the first developments, people realized that many of the schemes for (1.1) already known at the end of the 80s could fit into the fluctuation splitting formalism, at least when considering the steady limit of (1.1). Among these, cell-vertex FV schemes and the Streamline Dissipation (SD) FE scheme of Hughes and co-workers [146, 148, 147] (see [199, 201, DR07, 2, 95] for details) did fit in the framework. However, this approach allowed the development of new discretization principles, and of an original combination of some known techniques. *When focusing on the solution of the steady limit of (1.1), the most interesting contributions related to the fluctuation splitting approach are :*

**Genuinely Multidimensional Upwinding.** Problem (1.1) is known to admit solutions composed of waves traveling at speeds related to the eigenvalues of  $\vec{a}_\xi$  in (1.2). Some of these waves propagate in a precise direction, while other have an omni-directional propagation<sup>3</sup>. This behavior is typically taken into account in FV schemes by means of *upwind* numerical fluxes [215, 220, 259, 22]. However, in the FV context, upwinding only introduces a pseudo-one dimensional bias in the direction of mesh normals. The fluctuation splitting framework has led to the development of the so-called *multidimensional upwinding* (MU) principle [216, 217, 222, 237], that boils down to taking into account the multidimensional physical propagation directions when splitting the flux balance so that only downwind nodes (those not reached by the ink yet) receive some information. This can be done quite easily when (1.1) is a scalar equation, while, in the case of systems, it requires a decomposition into scalar problems, as we shall recall shortly. For steady state scalar problems, MU schemes showed a much reduced numerical dissipation (hence the name of “optimum” advection schemes) [217, 222, 237], a very fast iterative convergence to steady state [199, 201, 235, 152, 236, 258], and improved stability (e.g. with respect to the SD finite element scheme) and accuracy (w.r.t. FV schemes) on very distorted meshes [199, 201, 1].

**Linearity preservation** is normally referred to as the ability of a discretization to preserve piecewise linear solutions. Linearity preserving schemes are second order accurate. In the RD context it was soon recognized that these schemes are characterized by splittings that can be written as [216, 217, 178, 199, 201, 178, 241] (cf. figure 1.1 for the notation)

$$\phi_i^K = \beta_i^K \phi^K$$

where the *distribution coefficient*  $\beta_i^K$  is bounded w.r.t.  $\phi^K$ . This criterion has been formally related to the consistency of the scheme and to its truncation error quite late [1], showing its equivalence to formal second order of accuracy.

---

<sup>3</sup>typical examples in free surface flows are the propagation of a tracer (ink), and the ripples made by a stone thrown in a pond. Ink is carried with the flow and moves with the local water velocity, while the mass of water moved by the stone creates circular waves propagating in all the directions

**Positivity.** The theory of positive coefficient schemes (or positivity theory for short) is known to be a better candidate than TVD theory to construct high order schemes yielding non-oscillatory solutions [118, 238]. Positivity theory is the underpinning framework for the construction of shock capturing RD schemes. Since the first developments, positivity is ensured in the RD context by requiring that in every cell  $K$

$$\phi_i^K = c_{ij}(u_i - u_j) + c_{ik}(u_i - u_k) \quad \text{with} \quad c_{ij}, c_{ik} \geq 0$$

While some existing schemes could be recast into this framework (e.g. the first order cell-vertex FV scheme [199, 201]), a number of new positive schemes have been constructed [216, 217, 199, 201, 178]. Among these the multidimensional upwind N-scheme has been shown to be optimal in the sense that it has the minimal numerical dissipation among first order schemes [217, 222].

**Use of limiters.** One of the most original developments in the RD framework is the way in which FV limiter functions are used. Instead of limiting slopes or flux differences, limiters are used to obtain bounded distribution coefficients starting from a positive scheme. The basic idea is, given a limiter function  $\psi(\cdot)$ , to define [216, 217, 178, 199, 201, 178, 272, 241]

$$\beta_i^K = \psi(r_i) \quad \text{with} \quad r_i = \frac{c_{ij}(u_i - u_j) + c_{ik}(u_i - u_k)}{\phi^K} \quad \text{and} \quad c_{ij}, c_{ik} \geq 0$$

Under some standard conditions on the limiter, which we shall recall in part I, this construction leads to positivity preserving and second order discretizations. Different schemes are obtained by changing the limiter and the first order positive scheme. Note that this is precisely the opposite of what happens in FV schemes : the limiter is used here to pass from first to second order, and not to go back to first order (across discontinuities) as in FV.

Examples including limited RD variant of FV schemes are given in [199, 201] (see also [2] for a more recent review). The most successful of these schemes was known as the PSI scheme [216, 217, 241, 239, 199, 201, 1] obtained by limiting the multidimensional upwind N scheme. The PSI scheme has an accuracy very close to second order, fast iterative convergence, and remarkable stability even on very distorted unstructured grids [199, 201].

**Multidimensional Roe linearization.** The nonlinear character of (1.1) requires the satisfaction of a discrete conservation principle that guarantees that across discontinuities the discrete equations are consistent with the Rankine-Hugoniot jump conditions [231]. In the fluctuation splitting framework this was achieved by introducing a *multidimensional conservative linearization* of the flux Jacobian (1.2). The principle is exactly the same discussed in Roe's paper [215] discussing the one dimensional case, and it boils down to finding a parameter  $\mathbf{z}$  such that its average on the cell<sup>4</sup> verifies [86]

$$|K| \bar{\mathbf{a}}(\bar{\mathbf{z}}) \cdot \nabla \mathbf{z} \Big|_K = \oint_{\partial K} \mathcal{F} \cdot \hat{\mathbf{n}} \, dl \quad (1.3)$$

Note that for a linear interpolation over a triangular mesh cell, the gradient of  $\mathbf{z}$  on a cell is uniquely defined. For a number of simple problems, including the Euler equations for a perfect gas, this can be done easily. Not surprisingly, in the case of the Euler equations one finds out that  $\mathbf{z}$  coincides with the standard Roe parameter [86, 215].

---

<sup>4</sup>here by average we mean the simple arithmetic average of the values of  $\mathbf{z}$  in the nodes of the cell

**Hyperbolic and Hyperbolic/Elliptic decompositions** The effects of multidimensional upwinding can be further enhanced in the discretization of the steady Euler equations of gas dynamics by decomposing the system in its scalar hyperbolic sub-components, in the supersonic case, and in its hyperbolic and elliptic sub-components (the latter constituting a coupled Cauchy-Riemann type system) [199, 202, 36, 35, 136] in sub-sonic regions. These decompositions heavily rely on the use of the quasi-linear form of the system. The existence of a conservative linearization, at least for the Euler equations for a perfect gas, allowed the use of the quasi-linear form, while guaranteeing discrete conservation. With this technique, high order nonlinear multidimensional upwind schemes can be used on all the hyperbolic components, and ad-hoc central schemes can be used on the elliptic sub-system. The resulting discretization has remarkably low spurious entropy production even in presence of stagnation points [199, 201, 229], and can be even further enhanced to recover potential flow solutions in stagnation areas [210].

In summary, after roughly ten years of developments since Roe's initial idea, the fluctuation splitting framework had provided a truly multidimensional upwind second order discretization that had accuracy similar to that of the streamline upwind finite element scheme [55], while retaining some typical FV ideas such as positivity of the discretization and the use of limiters, even though in a completely new fashion. The very promising results obtained for typical steady inviscid aerodynamic configurations attracted some attention, and several applications of the methodology to different problems appeared, ranging from Magneto-Hydrodynamics [16, 82] (and later [17, 18, 225]), to hypersonics [88, 272], to turbulent flow predictions [54, 228, 179], to incompressible flow [33] (and later [34, 35, 269, 268]), to ion transport in electrochemical reactions [41, 42], to turbomachinery and industrial flows [135, 275], to ALE simulations and fluid structure interaction [181, 94, 95, 98, 97], to shallow water flows [113, 141, 205, 45] and others.

Direct comparisons with FV schemes [273] (and later [83, 124] and [1]) have shown generally that the accuracy of RD schemes is superior. All the discontinuity capturing features of the FV approach could be reproduced in the RD framework, thus making the method more appealing than the streamline dissipation FE scheme with artificial viscosity [55, 201, 202].

This is what motivated the further development of the schemes in the directions that will be discussed in the following paragraph.

### 1.1.3 Limitations

In the time spent as a member of the short training program of the von Karman Institute, I stumbled on some of the most important limitations of RD. At first, I was asked to work on explicit RD schemes for time dependent conservation laws [RD99, CdSR<sup>+</sup>00]. I realized there was no (or very little) understanding of how to apply this method to unsteady problems, or at least not while retaining all the nice properties they had for steady state simulations. Later on, I was asked to apply RD to a two-phase two-fluid model [RD00, VRD00] (see the more recent re-write [250]), and I had to face again the problem of dealing with time dependent flows in presence of source terms, and for models lacking a conservative linearization.

In 1999 RD methods were not mature enough to properly handle general time dependent conservation laws. The limitations were of course known and motivated the research on the following topics.

**Time dependent formulations.** The framework “compute balance-split-evolve” in the form used for steady computations (cf. figure 1.1), while appealing for its immediate heuristic physical interpretation, is not fully consistent for time dependent problems. By not fully consistent we mean that, whatever the time integration strategy, the scheme has a discretization error bounded by  $\mathcal{O}(h)$ , if  $h$  is the mesh parameter : the method is, in general, first order *in space*. Some bits of the discrete equations are missing for it to be of the correct order of accuracy.

Three basic approaches have been studied to recover second order of accuracy in time dependent problems

1. The earliest attempt was done at the VKI by means of a Petrov-Galerkin analogy in which the steady state RD equations are recast as a finite element scheme, formally similar to the SD of Hughes and collaborators [174, 107] (and later [9, 96, 97, 98, 95]). When applying the finite element statement in the time dependent case, a mass matrix appears, in equations : this mass matrix depends on the distribution coefficients  $\beta_i^K$  (cf. figure 1.1). The resulting discretization is indeed second order in space ;
2. Independently, at the University of Lund, Doru Caraeni argued that the relevant residual in the time dependent case is the whole equation, including the time derivative. He wrote a dual time stepping scheme in which the same compute balance-split-evolve is used, only including the time derivative in the elemental balance [54] (and later [53, 223, 136]). The resulting discretization features a mass matrix, however different than that obtained with the Petrov-Galerkin analogy. The resulting scheme is however indeed second order in space ;
3. The third approach investigated is quite different, however not quite original : the idea is to recast in a RD formalism the second order Taylor Galerkin scheme [99, 100] with lumping of the Galerkin mass matrix [145, RD99, CdSR<sup>+</sup>00] (and later [91, 144]). The scheme, which is referred to as the Lax-Wendroff scheme in most of the RD bibliography [199, 201, 145], is indeed second order in space and time.

These developments, of which more details are given in part I, give linear second order discretizations. However, a general framework to construct positive high order schemes is still missing, and in all the above references some form of nonlinear flux correction [39, 38, 40, 282] is applied to deal with discontinuities.

It is only much later that a technique allowing a formal construction of second (and higher) order positive schemes for time dependent flows has been proposed. This framework, bear similarities with both VKI’s and Caraeni’s approaches and has been proposed independently by R. Abgrall and his co-workers [9, 4] and by the group of the von Karman Institute, including this manuscript’s author. These developments are discussed in detail in part I.

**Conservation.** The use multidimensional upwind schemes relies on the existence of a (reasonably simple) conservative linearization of the multidimensional flux Jacobian over the mesh cell. Such a linearization is known in an exact (and simple) closed form only for some systems of conservation laws, and for linear interpolation on triangular cells. For systems not admitting such a linearization, conservative corrections have been suggested in [141, 81, 82].

The actual constraint for the discretization to provide the correct description of discontinuities is contained in the hypotheses of the Lax-Wendroff theorem proved in [8] (see also [11]) : the constraint is that in every cell, the split residuals sum up to the contour integral an *edge continuous* discrete flux. This discrete flux has always been thought of as  $\mathcal{F}(u_h)$  is  $u_h$

denotes the underlying continuous interpolation of the unknown on the mesh. In this spirit, the most advanced result, trying to cure this limitation, is that proposed in [6]. Given the continuous flux  $\mathcal{F}(u_h)$ , in the reference it is proposed to approximate the cell integral of its divergence as (cf. equation (1.2))

$$\int_K \nabla \cdot \mathcal{F}(u_h) = \int_K \tilde{a}(u_h) \cdot \nabla u_h \approx \sum_q \omega_q \tilde{a}(u_h(x_q)) \cdot \nabla u_h(x_q)$$

the  $\omega_q$  and  $x_q$  denoting the weight and the coordinates of a quadrature formula. The idea of the paper is to control the conservation error (*viz.* the quadrature error) such that it is always bounded by the truncation error. This boils down to adaptively switch from a very accurate quadrature across discontinuities to a low order integration formula in smooth areas [6]. Every quadrature point gives one quasilinear form on which one can apply any MU scheme. Ultimately the approach gives a set of discrete schemes, whose combination with the quadrature weights yields a conservative approximation. The authors provide formal proofs and sufficient numerical evidence (on linear triangles and for the Euler equations for a perfect gas) that their idea works. The only drawback is that, due to the poor behavior of approximate integration in correspondence of shocks, the method is extremely expensive.

**Stability and upwinding.** The implications of multidimensional upwinding on the stability of the discretization are poorly understood. Some understanding has come quite late in the development of the schemes [20, 5, 11, RVAD05, DR07], but it remains limited. Two attempts at giving a justification of the stability brought by MU have been made, in [11] by showing how MU influences the coercive character of a variational formulation of the schemes, and [2] by showing heuristically its influence on the algebraic well-posedness of the discrete problem (which indeed is also related to the coercivity issue).

More theoretical results concerning the PSI scheme are discussed in the recent papers [211, 56, 27]. In the papers the authors analyze a variational formulation similar to that studied in [11]. They are able to show, for linear triangles, existence and (quasi-)uniqueness of the discrete solution. However in these references : firstly, the authors do not consider the fully hyperbolic case and focus on advection-diffusion using a standard  $P^1$  continuous Galerkin discrete operator for diffusion ; second, in their proofs positivity and not multidimensional upwinding plays an important role in ensuring semi-positive definiteness of the bi-linear form ; lastly, the role of positivity is really more that of not destroying the coercive character of the  $P^1$  discrete Galerkin diffusion operator, the authors themselves expressing serious doubts on the positive-definiteness of the nonlinear advection operator itself [211] ;

**Formal understanding, variational setting, higher orders.** As the discussion above shows, one of the greatest handicaps for the development of the RD method is the very limited formal understanding available. Contrary to other techniques, there is neither a variational formulation, nor a proper functional space setting allowing to draw a perimeter within which solutions are sought. As it will be explained in part I, when starting from the initial multidimensional upwind schemes of Roe, one can deduce very few formal constraints for their generalization. This is possibly what has led several researchers to turn to more *easy to handle* techniques<sup>5</sup>. However, this is also what gives t the subject its charm, making it more challenging and exciting.

---

<sup>5</sup>such as Bram van Leer and Tim Barth whose contributions to the development stopped after few publications, *Alas*

The lack of a proper variational and functional setting translates in a lack of stability analysis (hence convergence and sharp error estimate) tools. This is also one of the major elements that has slowed down enormously the extension to higher orders, and other important developments such as the correct treatment of higher order derivatives, and the development of estimates for goal oriented mesh or polynomial adaptation.

#### 1.1.4 The appeal of the residual approach

With all the limitations mentioned above, the Residual Distribution framework had shown in its first ten years of developments enough potential to keep the scientific community interested in its developments.

The properties which have brought new players and fresh ideas into the development<sup>6</sup> are mainly related to the linearity preservation condition, and to the way in which discontinuity capturing is achieved :

**The residual property.** In this manuscript, the expressions *residual property*, *residual approach*, or *residual based* scheme and other similar ones indicate the simple fact that every operator in the discretization is expressed as an integral of the entire equation over a space (and eventually time) set within which the approximate solution and its derivatives are continuous, so that if the approximate solution were to be replaced by the exact one, the result would be the identity zero equals zero.

*If the method could be recast properly in a variational setting, the residual property would become the orthogonality principle.*

This residual property is what the RD method shared since its inception with stabilized continuous finite element schemes such as the streamline dissipation scheme of Hughes *et al.* [55] and the Taylor Galerkin scheme. Formally, it has been related to an integral truncation error for the first time in [1], justifying the experimental fact that linearity preserving schemes are second order. The simple rule distribute a local flux balance via bounded coefficients has become the paradigm for constructing (formally) high order schemes which has attracted several researchers and found application in other methods as well. Apart from the numerous developments published by R. Abgrall and co-workers (including this manuscript's author), we may mention its application to construct higher order schemes for relativistic hydrodynamics in [225], and its application to construct WENO finite difference schemes for non-smooth meshes [61, 62]. A very similar principle is at the basis of the RBC schemes of Lerat and Corre [167, 168, 76, 75, 73, 74, 77].

**Nonlinear schemes and shock capturing.** Since their earliest developments, Residual Distribution schemes embedded some form of discontinuity capturing. Differently from what was done in the finite element context this did not come in the form of an artificial viscosity [149, 201]. The multidimensional high order nonlinear limited RD schemes still remain unique in their attempt of retaining the same properties in all situations [216, 217, 201, 272]. In other words, nonlinear multidimensional upwind, positive, linearity preserving RD schemes remain multidimensional upwind, positive, and linearity preserving independently on the nature of the local solution. This is very different from what was done in every other technique and it is still today an appealing principle to avoid the suppression of smooth extrema when using discontinuity capturing.

---

<sup>6</sup>such as R. Abgrall and later C.-W. Shu and J.A. Rossmannith and others

**Multidimensional Upwinding and scalar decompositions.** For quite some time, residual distribution schemes were also known by the name of the multidimensional upwind schemes or rather *genuinely multidimensional upwind* schemes. Even though poorly understood on the mathematical side, the MU procedure proved a very powerful tool to achieve extremely fast iterative convergence in steady state calculations [235, 152, 236, 258], while allowing enhanced stability, and much lower numerical dissipation [202, 229, 210] on unstructured grids. Its applications to Magneto-Hydrodynamics [16, 17, 18, 82, 81] and to Shallow Water flows [113, 141, 205, 45] have shown a very high potential.

### 1.1.5 Relations with other techniques

As said in the beginning, there exist a large number of techniques to solve (1.1) numerically. Many of these share with residual distribution the aim : overcoming the accuracy limitations of finite volume schemes, while retaining some of their properties such as monotonicity of the solution. Not surprisingly, all these methods also share their “inception date” which amounts at roughly the end of the 80s when FV schemes had shown most of their potential and also started showing their limitations in the multidimensional case.

Fluctuation splitting was a re-formulation of one dimensional flux difference splitting and indeed they are the same thing in 1d. In the multidimensional case it is harder to make the link between the finite volume method and RD. One point of contact of the two is shown in [201, 2, 95], where cell-vertex FV schemes are recast in the RD formalism as a basis to construct new RD schemes. A more interesting point of view would be probably to exploit the 1d equivalence to come up with RD based FV fluxes. Philosophically, this would be similar to the singular residual technique used in [196, 197, 278].

The possibility of further combining the two techniques is investigated also in [61, 62] where a finite difference/residual distribution approach based on WENO interpolation and FV fluxes is proposed for steady state problems. As mentioned already, this is a typical example of how the residual property can be used to improve other discretization techniques.

Among the other emerging methods aiming at replacing the FV approach, a very successful one is the Discontinuous Galerkin (DG) one. After the seminal work of Reed and Hill on the neutron transport equation [212], the development of DG schemes for (1.1) was taken over by B. Cockburn and C.-W. Shu with a series of 5 papers that started in 1988 [69] and went on for a few years [70, 68, 66, 71]. Since then, many other successful adaptations of the approach to nonlinear hyperbolic problems have appeared (see [67] for a review). The DG method has probably become the most popular method to solve (1.1) numerically. Its success, is a lot related to ease of implementation, well established mathematical setting, and great flexibility. The enormous number of investigators has rapidly brought it to a high level of maturity.

Despite its many advantages, the DG method is not a flawless technique. One of the known flaws is its cost in terms of number of degrees of freedom [31, 150], associated to the locally discontinuous nature of the approximation. In addition, despite of its in-cell residual character, it relies on the same stabilization mechanisms of FV schemes (Riemann Fluxes on cell edges). This does have consequences on the accuracy of the scheme, for example, in presence of source terms. Shock capturing in the DG framework is also still a subject of intense research, and only recently a robust and accurate maximum principle satisfying and positivity preserving approach has emerged [283, 284, 285].

Bridges between the DG and the RD approach have been discussed already by several authors [12, 140]. The simplest idea is to replace the in cell residual component of the DG method by a RD scheme. More complex constructions are also possible. The advantage of such a hybrid approach would be to incorporate the shock capturing features and the stabilizing multidimensional upwinding effects into the DG framework, or, to exploit the flexibility of DG's discontinuous approximation in the RD context.

There is plenty of other methods for (1.1). We already mentioned the RBC approach of Lerat and Corre [167, 168, 76, 75, 73, 74, 77], very similar to the residual distribution method, however mainly developed for structured meshes. An interesting family of schemes is related to the so-called ADER approach [249, 226, 103, 104], which is a generalization of the simple Lax-Wendroff scheme. In itself ADER is not a class of spatial discretizations but rather a paradigm to devise coupled space-time discretizations. Its appeal is precisely that it does not require a high order time discretization scheme, which is replaced by the coupling of a Taylor series expansion in time with application of the Cauchy-Kovalewski procedure. Some issues such as shock capturing remain however similar to those of the underlying spatial discretization. Other relatively successful techniques include the Spectral Finite volume/difference approach of Z.J.Wang (the first four installments in [263, 264, 265, 267]). The approach of Wang is basically a FV approach, except that in-cell polynomial approximations are constructed by solving by several sub-cell averages used to obtain the local polynomial expansion. The method has shown very high accuracy and is based on a very interesting idea. What makes it complex is the sub-cell subdivision which has enormous impact on the accuracy and especially stability of the polynomial interpolation [252, 251, 253, 254]. For high order polynomials the sub-division for a given polynomial degree is not unique, it is generally geometrically very complex, and many of the known partitions suffer from lack of stability [252, 251].

As it is (hopefully) apparent from this short and far from complete overview, even though several techniques are available to solve (1.1) numerically, there is space for improvements both within each class of methods and in general.

Some ideas seem to emerge as having a general interest, such as : the use of sub-cell high order polynomials to reach higher accuracy ; the quest for a positivity preserving approach ; some form of residual character. These are the principles that are exploited in the developments recalled in the next paragraph.

### 1.1.6 My contributions

My work on the development of high order residual distribution schemes can be grouped into 5 categories : contribution to the *formal understanding* of the schemes ; derivation of a *general conservative framework* ; construction of *high order schemes for steady problems* ; construction of *high order schemes for time dependent problems* ; *applications*.

**Formal understanding.** These contributions concern the positivity, the stability (energy analysis), and the error analysis. In particular, following the analysis of FV schemes of [22], a structured, abstract, algebraic description of the schemes, allowing their analysis in terms of the so-called Local Extremum Diminishing (LED) principle, and allowing their subsequent positivity and discrete maximum principle analysis for different explicit and implicit time-stepping choices, including problems with source terms, has been provided [DR07, MRAD03, RD02].



For linear scalar problems, the energy stability of the schemes has been related to the above mentioned algebraic properties to find back the result that LED schemes all LED schemes are energy stable, as well as that fully discrete stability depends on the time integration strategy [DR07]. The energy analysis of the most successful MU schemes, the LDA and N, shows that they can be written as a local (cell) net energy balance plus a dissipative term [DR07]. Moreover, it is shown that all Multidimensional Upwind schemes are strictly dissipative in cells with only one downstream node [DR07, RVAD05].

The error analysis of [1, 11] has been generalized to non-homogeneous problems [RVAD05, RAD07, DR07], and to time dependent problems with general implicit or explicit multi-step discretizations in time [RAD07, RA10], which include Galerkin type time discretizations, thus encompassing space-time schemes as well. Lastly, a framework for adjoint error analysis is under development to allow goal oriented grid refinement [DRAD11]. The reader interested can consult the last reference for details on this last topic which has been left out of this manuscript.

**General conservative framework.** A general conservative framework that has substantially simplified the discretization, and replaced almost completely the use of Roe's linearization, has been proposed and tested on different element types and different complex conservation laws [CRD02, QRCD02, RCD04, RCD05, DR07]. This paradigm is based on the simple idea that the choice of the discrete flux is independent on that of the solution, albeit verifying all the hypotheses for convergence to the correct weak solutions (continuity) and accuracy (degree of polynomial approximation larger or equal that that of the solution). The approach does not prevent the use of scalar decompositions for the system [124]. This paradigm is currently the one used in all the higher order ( $> 2$ ) schemes, for both steady and time dependent computations.

**High order schemes for steady problems.** Substantial contributions were given to the development of very high order ( $> 2$ ) versions of RD. Two approaches are followed : one based on a centered positivity preserving distribution [ARN<sup>+</sup>06, ARTL07, ALR08b, ALR08b, LAR09, ALRT09, RAA09, ALR10, ABJR11, ALR11], the other trying to retain the multidimensional upwinding property [RVAD05, VRD06, ARN<sup>+</sup>06, RVAD08, VQRD11]. The first approach has shown promising features concerning monotonicity and accuracy. The MU schemes, have much faster convergence to steady state (for the same accuracy). However, they still have some monotonicity problems and are overly complex.

**High order schemes for unsteady problems.** I contributed to the conception and analysis of implicit [MRAD03, RCD04, RCD05, RA06, RB09b], explicit [CdSR<sup>+</sup>00, RA10], and space-time [CRDP01, DRD02, CRD03b, RAD03, DRD03a, DRD03c, CRD03a, DRD05, DR07, HR09, HR10, HR11, HRS11] residual distribution schemes for time dependent conservation laws. My work has brought sufficient knowledge to construct high order genuinely explicit schemes, as well as unconditionally (w.r.t. the time step) positive ones. While improvements are continuously studied, after almost 12 years (cf. beginning section §1.1.3) RD can be used to solve arbitrary time dependent conservation laws with, at least, second order of accuracy.

**Applications.** Some of the developments have been pushed forward or tailored to applications such as two-phase flow [VRD00, CRD03a, RRWD03, RCD04, SFW<sup>+</sup>05, RCD05], Magneto-Hydrodynamics [CdSR<sup>+</sup>00, CRD02], and free surface flows [RAD07, DR07, RB09b,

RB09a, CBR<sup>+</sup>09, Ric09a, Ric09b, Ric11, HRS11]. Shallow Water flows, in particular, have been and are still a perfect playground to further improve and test the positivity preserving and residual principles underpinning the RD schemes I develop. Next section and part II of this manuscript are entirely devoted to this topic.

It must be remarked that these developments have

- *great importance* (of course) in the niche of the developers of the Residual Distribution techniques ;
- *great interest* in the community of developers of high order residual based discretizations, in terms of how far the residual property can be pushed, and what are the other important requirements when solving (1.1) ;
- *some impact* on the community of developers of high order schemes for hyperbolic conservation laws, in terms of understanding of what the best guiding principles are in designing schemes for (1.1) ;

Most of this work would never have been possible without the collaboration of my colleagues. I am (especially) indebted for their contributions to my past and current work on residual distribution to

- Cedric Tavé, Robin Huart, Guillaume Baurain, Pascal Jacq, Dante DeSantis, and of course Rémi Abgrall at INRIA ;
- Nadege Villedieu, Stefano D’angelo, Martin Vymazal, Arpaad Csik, Jirka Dobes, Tiago Quintino (and many others), and of course Herman Deconinck at VKI ;
- Andrzej Warzynski, Domokos Sarmany and of course Matthew Hubbard of the School of Computing at Leeds University ;
- Adam Larat at École Centrale Paris ;

## 1.2 High order schemes for free surface flows

### 1.2.1 Problem setting

Free surface flows are relevant in a large number of applications, especially in civil and coastal engineering. The problems concerned are either (relatively) local, such as dam breaks and flooding, overland flows due to rainfall, nearshore wave propagation and interaction with complex bathymetries/structures, and tidal waves in rivers, or global such as in ocean or sea basin models for the study of *e.g.* tsunami generation and propagation.

The simulation of such flows can be carried by solving directly the three dimensional Navier-Stokes equations. However, for many applications, including *e.g.* nearshore wave propagation and flooding, simplified models obtained by combining vertical averaging and some form of thin layer approximation provide reliable results. The applicability of such models depends on the nature of the flow and on the hypotheses at their basis [165, 37].

The simplest among these models is the so-called Shallow Water model. The model assumes that the waves developing in the flow are *long* (small ratio amplitude/wavelength), and of a hydrostatic vertical variation of the pressure [117, 175] .

The first order approximation (in terms of the ratio amplitude/wavelength) equations constitute a non-homogeneous hyperbolic system where the effects of the variation of the bathymetry and the viscous friction on the bottom are modeled by the source terms [117, 175]. More complex nonlinear models can be obtained, by including higher order terms, and depending on the hypotheses on the flow [117, 175, 165, 37].

Part II of this manuscript, discusses my work on application and the further development of residual distribution for the solution of the Shallow Water system that reads

$$\begin{aligned} \partial_t d + \nabla \cdot (d\vec{v}) + R(x, y, t) &= 0 \\ \partial_t (d\vec{v}) + \nabla \cdot (d\vec{v} \otimes \vec{v} + p(d)\mathbf{I}) + gd(\nabla b + k_f \vec{v}) &= 0 \end{aligned} \quad (1.4)$$

where  $d$  represents the depth,  $\vec{v}$  the (vertically averaged) local velocity,  $R$  is a source of mass (*e.g.* associated to rainfall),  $b$  is the bathymetry,  $k_f$  is a friction coefficient, and the hydrostatic pressure is given by

$$p(d) = g\frac{d^2}{2}$$

System (1.4) is endowed with a mathematical entropy coinciding with the total energy [245, 246, 130, 134], it is hyperbolic, and characterized by the physical constraint of the non-negativity of the depth.

The amount of literature related to the solution of (1.4) is extremely vast. This model finds applications in oceanography, hydrology, and meteorology (see *e.g.* [242, 142, 51, 112, 243, 244] and references therein). The main challenges when solving (1.4) numerically are related to the discretization of the bathymetry and friction terms, and to the numerical treatment of nearly dry regions ( $d = 0$ ). For the first issue, one speaks often *asymptotic preserving* character or *well balancedness* of a discretization. The second issue is what is referred to as the wetting/drying strategy.

### 1.2.2 Well balancing

The asymptotic preserving nature of a scheme is related to its behavior when some parameter is very large (or very small..). The asymptotic preserving behavior is related *e.g.* to the long time behavior of the discrete solution, or, equivalently, to the behavior of the scheme when *e.g.* the friction coefficient in (1.4) becomes large. It is a measure of how the scheme handles the equilibrium between the different terms, when some source of stiffness is introduced. In this cases, an asymptotic analysis of the original equations can be used to infer what the *physical asymptotic* behavior of the solution is. If a similar analysis can be done on the numerical scheme, showing that it does reduce to a consistent approximation of the asymptotic equations, then the scheme is asymptotic preserving [154]. There is plenty of literature on various forms and applications of schemes preserving some asymptotic behavior (see *e.g.* [154, 122, 102, 166] and references therein for an overview).

Well balancing, instead, refers to the ability of the discretization to preserve exactly<sup>7</sup> some steady equilibria involving the existence of a set of invariants. The typical example is the so called *lake at rest state* involving a flat still free surface, that should be remaining flat whatever the shape of the bottom. This property is what one refers to as *Conservation property*, or *C-property* [26] or well-balancedness [121]. It becomes important when one is

<sup>7</sup>or within some mesh size dependent bounds, usually more favorable than the accuracy of the scheme

interested in flows that, at least locally, are perturbation of one of these steady equilibria, so that numerical perturbations might interfere with the actual flow giving wrong results. There is plenty of literature discussing several different approaches to the preservation of steady equilibria for (1.4), in particular the lake at rest state. Most of these developments have taken place in the Finite Volume community, and are thought in terms of one-dimensional flows (see *e.g.* [26, 121, 111, 196] and references therein). The basic approach boils down either to the inclusion of a source term contribution in the FV numerical flux, so that the correct equilibrium is found at the discrete level [26, 121, 143], or to the rewriting of the system in a relaxation form, where an appropriate integral of the source term is added to the physical flux in the Maxwellian on the right hand side [90, 227]. The extension to multiple space dimension is often done in a dimension by dimension basis on structured grids (see [278, 197, 196, 195], for recent examples), or introducing local one dimensional problems along some geometrical directions (*e.g.* normals to grid faces) [143, 89, 189, 19]. These modified FV fluxes are also used in the context of discontinuous Galerkin schemes to retain the C-property (see *e.g.* [276, 106]). Exceptions to this rule are the wave propagation scheme of LeVeque [171], continuous finite elements discretizations as the least squares Galerkin approach of G.Hauke [130], and Residual Distribution schemes [45].

### 1.2.3 Positivity preservation and wetting/drying

The computational treatment of nearly dry areas<sup>8</sup>, meaning with a water depth  $d$  very small however positive, involves the solution of the following issues : ensuring that in these regions no unphysical negative depths are obtained ; handling some ill-posed problems such as the computation of the local velocity given depth  $d$  and discharge  $d\bar{v}$  ; preserving the well balanced character of the method when  $d \ll 1$ .

These three issues are not independent and the large majority of the wetting/drying treatments discussed in literature boil down to : rely on the use of some positivity preserving scheme to be able to keep the depth non-negative ; introducing a cut-off of some sort on the velocity (and mass flux) to avoid zero over zero type divisions ; modify the *numerical* slope of the bathymetry used in the discrete equations ; employ an implicit (split or unsplit) treatment of the friction term to handle the stiffness associated to this term in dry areas<sup>9</sup>. These ideas can be put in practice in various ways, depending on the initial formulation of the method, on the techniques used to reach higher order of accuracy, and on the type of nonlinear mechanism used to combine high order and preservation of the positivity. For an overview see [47, 46, 57, 58, 19, 106, 189, 279, 277] and references therein.

### 1.2.4 My contributions

My work finds his motivation in the attempt of dealing with the issues discussed above in a truly multidimensional setting. The residual distribution approach gives this setting.

The earliest work on the subject has been done by M. Hubbard in his PhD. In particular, in [113] Hubbard and Garcia-Navarro present a non conservative discretization based on multidimensional upwind RD schemes and scalar wave decompositions. The advantages related to the use of a discretization based on triangles would seem to make the schemes competitive with those currently in use. Similar results have been shown independently in [205]. Later

---

<sup>8</sup>obviously completely dry areas do not pose problems, the equations reducing to the identity  $0 = 0$

<sup>9</sup>It actually depends on the friction model chosen, however most physical models tend to yield coefficients that are unbounded for  $d \rightarrow 0$ , leading to zero velocity at the front

on, in [141], Hubbard and Baines proposed a correction rendering the discretization mass and momentum conserving. However, while the underlying scheme is still multidimensional upwind, the correction is treated along the source terms in a centered way. Lastly, in [45] Brufau and Garcia-Navarro extended the initial work of [113] studying in some more detail the issue of well-balancing and of wetting/drying.

The above developments have shown interesting results for problems relevant to hydraulic engineering. The work of [45] has also shown (citing from the abstract) “the necessity of a multidimensional upwind discretization of the source terms” in order to achieve well balancing on unstructured meshes with this approach. Some bricks are, however, still missing. My contribution to the development has been the following.

**Conservative high order approach for time dependent flows.** Taking advantage of my work on conservative formulations of the method, a genuinely conservative formulation of RD for shallow water flows simulations on unstructured grids has been proposed [RAD07, DR07], allowing to overcome and understand the limitations of previous work [141].

Similarly, the development of high order accurate formulations for time dependent flows has permitted the construction of truly second order unstructured grid schemes for shallow water flows simulations on unstructured grids [Ric09b, RB09b, RB09a, HRS11].

**Source terms and general framework for C-property.** The understanding of the asymptotic accuracy of RD via integral truncation error analyses has given the tool to properly include the source terms. In particular, the residual property has been exploited to give a general rule for the satisfaction of the C-property, independently on the upwind nature of the distribution [Ric09a, Ric11, HRS11].

**Multidimensional framework for generalizations of C-property** In the residual framework the generalization of the C-property to equilibria less trivial than the lake at rest is easier. In particular, because of the nature of the schemes, this is obtained directly in a multidimensional setting and on unstructured grids [Ric09a, Ric11]. These non-trivial equilibria include constant total energy flows, and flows in sloping channels with friction.

**Multidimensional wetting/drying and positivity.** These generalizations have been coupled with a wetting/drying methodology that, as in [45], modify the numerical slope of the bathymetry. This approach has been coupled with a positivity preserving discretization to allow the simulation of real applications including long wave run-up, flooding and overtopping on unstructured grids [RB09b, RB09a, Ric09b].

In particular, the positivity condition imposes, for most standard time stepping approaches, a time step limitation [32, 119]. Similar limitations are found for positivity preserving RD schemes for time dependent flows [DR07, RB09b]. In the framework of residual methods, these limitations can be rendered the scheme inefficient, due to the presence of a (often solution dependent) mass matrix [RAD07, RB09b]. Solutions to these limitations are proposed by devising on one hand genuinely explicit formulations that preserve positivity under a time step constraint, unconditionally positive space-time schemes [RA10, HRS11].

Remark that these contributions

- Have brought the RD methodology to at least the same level of maturity as state of

the art FV schemes, for the simulation of shallow water flows on unstructured grids [46, 19, 189] ;

- provide a *contribution to the understanding* of balanced discretizations for free surface flows, giving some *general principles* which can be used in the context of other methods ;
- Give a very competitive alternative to standard FV approaches for the simulation of free surface flows.

I am of course indebted for these developments to my collaborators, in particular to

1. Herman Deconinck at the von Karman Institute ;
2. Rémi Abgrall at INRIA ;
3. Andreas Bollerman during his master thesis and his PhD at RWTH, University of Aachen ;
4. Domos Sarmany and Matthew Hubbard of the School of Computing at Leeds University.



## Part I

# Residual discretizations for hyperbolic conservation laws





This part has to give sufficient background and references for the reader to understand the contributions made in the field of residual distribution schemes.

These contributions have been listed in the introductory chapter. Some of them will be discussed in some detail here, starting again from their context and motivation and arriving to the results obtained.

It must be remarked that these developments have

- *great importance* (of course) in the niche of the developers of the Residual Distribution techniques ;
- *great interest* in the community of developers of high order residual based discretizations, in terms of how far the residual property can be pushed, and what are the other important requirements when solving hyperbolic conservation laws ;
- *some impact* on the community of developers of high order schemes for hyperbolic conservation laws, in terms of understanding of what the best guiding principles are in designing schemes for hyperbolic conservation laws.



## Chapter 2

# Residual distribution schemes

### 2.1 Generalities and notation

This part of the manuscript is concerned with the numerical approximation of solutions to the system of nonlinear Partial Differential Equations (PDEs)

$$\partial_t u + \nabla \cdot \mathcal{F}(u) = 0 \quad \text{on} \quad \Omega \times [0, T_{\text{fin}}] \subset \mathbb{R}^d \times \mathbb{R}^+ \quad (2.1)$$

with  $u$  a set of conserved quantities, and  $\mathcal{F}$  a conservative flux. We consider the two-dimensional case  $d = 2$ , however the discussion generalizes trivially to the three space dimensions. The *conservation law* (2.1) is assumed hyperbolic and in particular it enjoys all the classical properties (see e.g. [231] and [21, 246, 245, 125, 184, 134]) such as existence of an entropy pair, symmetrizability, and that the projection of flux Jacobian (which in general is a  $d$ -dimensional vector of matrices) on the arbitrary  $\vec{\xi} \in \mathbb{R}^d$  direction

$$a_\xi = \vec{a}(u) \cdot \vec{\xi} = \partial_u \mathcal{F}(u) \cdot \xi \quad (2.2)$$

is diagonalizable with linear eigenvalues and linearly independent eigenvectors. Moreover, being nonlinear, (2.1) admits discontinuous solutions compatible with the entropy conditions, and with the Rankine-Hugoniot relations [231]. Problem (2.1) is also endowed by a set of boundary conditions  $F(u) = G$  weakly on the inflow part of  $\partial\Omega$ , and with an initial condition

$$u(x, y, z, t = 0) = u_0(x, y, z) \quad (2.3)$$

Many of the developments discussed, will consider the steady state limit of (2.1)

$$\nabla \cdot \mathcal{F}(u) = 0 \quad \text{on} \quad \Omega \subset \mathbb{R}^2 \quad (2.4)$$

Let  $\Omega_h$  be an unstructured tessellation of the two-dimensional spatial domain  $\Omega$ , composed of a set of non-overlapping elements  $K$  (cf. 2.1). With  $h$  we denote the reference mesh size (e.g. largest element diameter). Each element is endowed with a set of degrees of freedom identified with some nodes as shown on the right on figure 2.1. The set of nodes  $j \in K$  will be sometimes locally numbered as  $(1, 2, \dots, j_K)$ . The generic degree of freedom of the mesh will be instead referred to either as  $\sigma \in \Omega_h$ , or with small italic letters (e.g.  $i, j, k$  etc.).

We also introduce the set  $K_i$  of all the elements containing  $i$  as a degree of freedom. So that if  $K \in K_i$ , then  $i \in K$ . By abuse of notation, we will shall also say that  $j \in K_i$  is there

exist a  $K$  such that  $i, j \in K$ . Similarly, let  $f \in \Omega_h$  be the generic element face, and denote by  $F_i$  the set of faces containing  $i$  as a degree of freedom.

In the time dependent case, the temporal domain  $[0, T_{\text{fin}}]$  is broken in slabs  $[t^n, t^{n+1}]$  of width  $\Delta t^n = t^{n+1} - t^n$ . We set  $\Delta t = \max_n \Delta t^n$ .

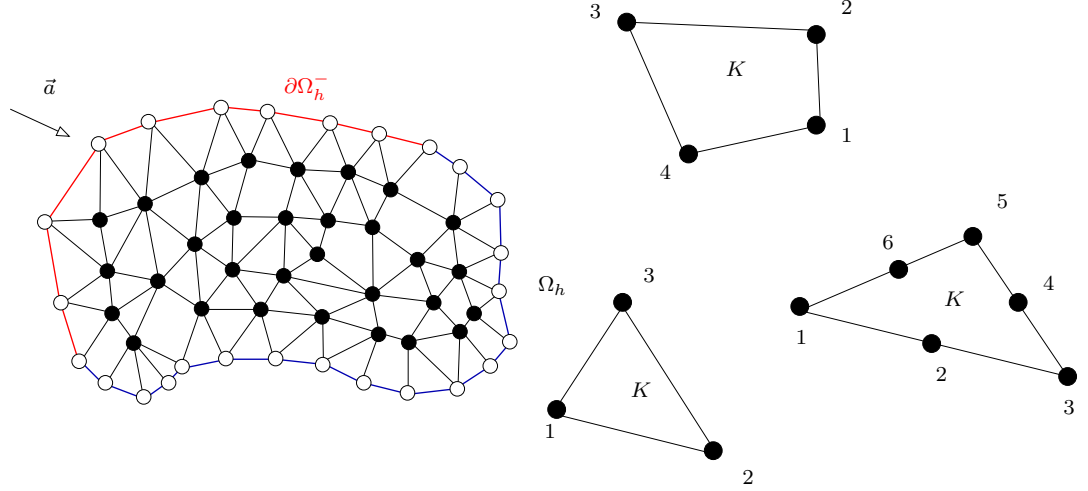


Figure 2.1: Mesh and elements

Given a set of values of a quantity  $w$  at the degrees of freedom, on each element  $K$  we consider the polynomial built interpolating these values :

$$w_h|_K = \sum_{j=1}^{j_K} w_j \varphi_j \quad (2.5)$$

where the basis functions  $\varphi_j$  are standard edge continuous Lagrange basis polynomials [100, 286] verifying

$$\begin{aligned} \varphi_i(x_j) &= \delta_{ij}, \quad \forall i, j \\ \sum_{j=1}^{j_K} \varphi_j &= 1 \\ \sum_{j=1}^{j_K} \nabla \varphi_j &= 0 \end{aligned} \quad (2.6)$$

**Remark 2.1.1** (Continuous approximation). *Unless otherwise stated, the approximation  $w_h$  is supposed to be continuous and global. In other words, the values  $\{w_j\}_{j=1}^{j_K}$  on the boundaries of  $K$  are uniquely and globally defined on  $\Omega_h$ . The overall continuous approximation is obtained as*

$$w_h = \sum_{K \in \Omega_h} w_h|_K$$

with  $w_h|_K$  given by (2.5). Note that, whenever no confusion is generated, the restriction  $|_K$  will be omitted to simplify the notation.

## 2.2 Fluctuation splitting on linear triangles

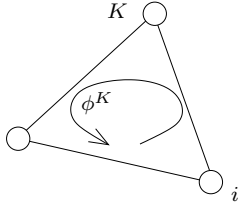
This section will provide a (very) basic understanding of the *fluctuation splitting/residual distribution* approach, as initially developed by P.L.Roe, H. Deconinck and co-workers. These schemes were developed in the simpler setting in which (2.4) reduced to the scalar advection equation

$$\vec{a} \cdot \nabla u = 0 \quad \text{on} \quad \Omega \subset \mathbb{R}^2 \quad (2.7)$$

where  $\nabla \cdot \vec{a} = 0$ , and with boundary conditions

$$\int_{\partial\Omega} (\vec{a} \cdot \hat{n})^- (u - g) = \int_{\partial\Omega^-} \vec{a} \cdot \hat{n} (u - g) = 0 \quad (2.8)$$

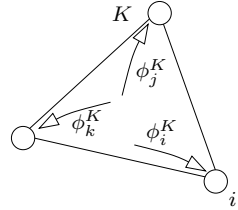
We seek a numerical approximation of the solution of (2.7) with boundary conditions (2.8), on a triangulation  $\Omega_h$  of the spatial domain. The initial idea we exploit to do this is due to Phil Roe, and it is a multidimensional generalization of his interpretation of flux difference splitting in the context of *fluctuations and signals* [215, 218, 219, 216, 217] :



Given initial values of the solution in the nodes of the mesh

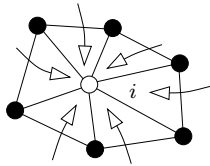
1.  $\forall$  triangles  $K$  compute the *fluctuation/residual*

$$\phi^K = \int_K \vec{a} \cdot \nabla u_h|_K \left( \approx - \int_K \partial_t u_h \right) \quad (2.9)$$



2.  $\forall$  triangles  $K$  distribute the fluctuation to the three nodes of  $K$ . Let  $\phi_j^K$  denote the *amount* of fluctuation sent to node  $j \in K$ , then the *conservation/consistency* requirement is

$$\sum_{j=1}^{j=j_K} \phi_j^K = \phi^K \quad (2.10)$$



3.  $\forall$  nodes  $i \in \Omega_h$  evolve the nodal values by assembling *signals* from surrounding triangles

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \phi_i^K \quad (2.11)$$

where the *dual cell*  $C_i$  is normally defined as the cell obtained joining cell centers with edge mid-points (cf. right on figure 2.2) and its area is

$$|C_i| = \sum_{K \in K_i} \frac{|K|}{3} \quad (2.12)$$

with  $|K|$  the area of element  $|K|$ . The key of the approach is of course the definition of the splitting at point 2.

**Remark 2.2.1** (Steady state). *The update (2.11) has to be considered here as an iterative means to reach a steady discrete solution, whose nodal values will actually satisfy the equations*

$$\sum_{K \in K_i} \phi_i^K = 0, \quad \forall i \in \Omega_h \quad (2.13)$$

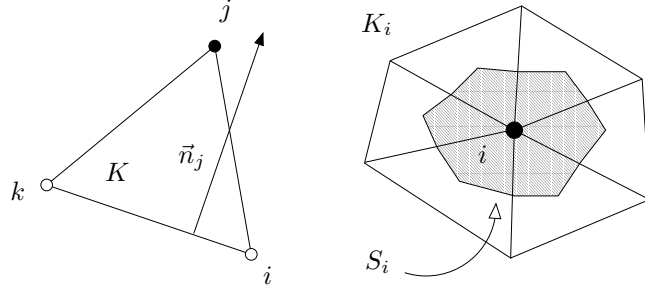


Figure 2.2: Median dual cell  $S_i$  and nodal normal  $\vec{n}_j$

### 2.2.1 Boundary conditions

Before discussing the distribution criteria, it is important to stress that, even though in most fluctuation splitting papers the method is described by (2.9)-(2.10)-(2.11), there is one big missing piece of information : the boundary conditions (BCs).

The most general way to introduce these conditions is to consider  $\forall f \in \partial\Omega_h$  the *face fluctuations*.

$$\phi^f = \int_f (g_h^* - u_h) \vec{a} \cdot \vec{n} \quad (2.14)$$

where, in order to take into account the compatibility condition implicit in (2.8), the boundary flux  $(g^* \vec{a}) \cdot \vec{n}$  has been introduced. The compatibility is easily ensured by taking *e.g.*

$$g^* = \frac{1 + \text{sign}(\vec{a} \cdot \vec{n})}{2} u + \frac{1 - \text{sign}(\vec{a} \cdot \vec{n})}{2} g$$

Face fluctuations can be split to the degrees of freedom  $j \in f$  by defining  $\phi_j^f$  distributed residuals such that

$$\sum_{j \in f} \phi_j^f = \phi^f \quad (2.15)$$

Finally, the complete discrete model reads

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \phi_i^K - \sum_{f \in F_i} \phi_i^f \quad (2.16)$$

For some problems, especially in the scalar case, BCs are strongly imposed by explicitly setting  $u_i^{n+1} = g(x_i)$ .

### 2.2.2 Positivity and discrete maximum principle

We start by reviewing some splitting principles. An important criterion is the *local positivity* of the distribution, which is related to positive coefficient theory which has replaced the TVD theory to construct high order schemes [118, 238, 22].

**Definition 2.2.2** (Positive scheme). *A (locally) positive scheme is one for which*

$$\phi_i^K = \sum_{\substack{j \in K \\ j \neq i}} c_{ij}(u_i - u_j), \quad c_{ij} \geq 0 \quad \forall j \in K \quad (2.17)$$

Positivity is the key to the construction of non-oscillatory schemes [216, 201] :

**Proposition 2.2.3** (Local Positivity and *discrete maximum principle*). *Locally positive schemes, combined with the evolution step (2.11) verify the discrete maximum principle*

$$\min_{j \in K_i} u_j^n \leq u_i^{n+1} \leq \max_{j \in K_i} u_j^n \quad \forall i \in \Omega_h$$

*under the time step limitation*

$$\Delta t \leq \min_{i \in \Omega_h} \left( \frac{|C_i|}{\sum_{K \in K_i} \sum_{\substack{j \in K \\ j \neq i}} c_{ij}} \right)$$

*Proof.* The proof follows from the positivity of the  $c_{ij}$ s and time step restriction, and from

$$u_i^{n+1} = \left( 1 - \frac{\Delta t}{|C_i|} \sum_{K \in K_i} \sum_{\substack{j \in K \\ j \neq i}} c_{ij} \right) u_i^n + \frac{\Delta t}{|C_i|} \sum_{\substack{j \in K_i \\ j \neq i}} \sum_{K \in K_i \cap K_j} c_{ij} u_j^n$$

□

### 2.2.3 Linearity preservation and accuracy

On linear triangles, the key to construct second order schemes is in the so-called *linearity preservation* property [216, 201]. A linearity preserving scheme is defined as follows.

**Definition 2.2.4** (Linearity preservation). *Let  $\{\beta_j^K\}_{j \in K}$  be a set of distribution coefficients uniformly bounded with respect to  $h$ ,  $u_h$ ,  $\phi^K$ , and with respect to the data of the problem ( $\vec{a}$ , boundary data, etc. etc.), and verifying the consistency property*

$$\sum_{j \in K} \beta_j^K = 1 \quad (2.18)$$

*A Linearity Preserving scheme is one for which*

$$\phi_i^K = \beta_i^K \phi^K \quad (2.19)$$

**Proposition 2.2.5** (Linearity preservation and second order of accuracy). *Linearity preserving schemes are second order accurate.*



The proof of this property allows to introduce one of the most important tools available for the construction of high order RD schemes : consistency and truncation error analysis. Even though the validity of proposition 2.2.5 has been known since the very beginning [216], this analysis has only been introduced in [1]. We shall recall the analysis here, following [11]. Its generalizations are discussed in the following chapters.

**Definition 2.2.6** (Truncation error and accuracy). *Let  $\psi$  be a compactly supported smooth function,  $\psi \in C_0^{r+1}(\Omega)$ . Let  $\Omega_h$  be an unstructured grid composed of non-overlapping elements. On the generic element  $K \in \Omega_h$  consider the  $r$ -th degree continuous Lagrange approximation (2.5). Let in particular  $\psi_h = \sum_{j \in K} \psi_j \varphi_j$  be the  $r$ -th degree polynomial approximation of type (2.5) of  $\psi$ , the values  $\psi_j$  being obtained by Galerkin projection. Consider now an exact smooth function  $u \in H^{r+1}$  verifying (2.7) in a classical sense :  $\vec{a} \cdot \nabla u = 0$  in  $\Omega$ . Let  $u_h$  be of polynomial approximation of degree  $r$  of type (2.5) of  $u$  obtained by Galerkin projection. Let now  $\phi_j^K(u_h)$  the value of the split residuals (2.10) obtained when replacing the nodal values of the solution obtained with the scheme by the values  $u_j$  of the Galerkin projection of  $u$ . We define the integral truncation error  $\epsilon(u_h, \psi)$*

$$\epsilon(u_h, \psi) = \sum_{j \in \Omega_h} \psi_j \sum_{K \in K_j} \phi_j^K(u_h) = \sum_{K \in \Omega_h} \sum_{j \in K} \psi_j \phi_j^K(u_h) \quad (2.20)$$

We shall say that a scheme is  $r+1$  order accurate if it verifies the truncation error estimate

$$|\epsilon(u_h, \psi)| \leq C(\Omega_h) h^{r+1}$$

The following general characterization is possible.

**Proposition 2.2.7** (Accuracy condition). *A sufficient condition for scheme (2.9)-(2.10)-(2.13) to be  $r+1$  order accurate in the sense of definition 2.2.6 is that*

$$|\phi_i^K(u_h)| \leq C(\Omega_h) h^{r+2} \quad \forall K \in \Omega_h, \quad \forall i \in K \quad (2.21)$$

*Proof.* With the notation of definition 2.2.6, we start by introducing the *Galerkin* residuals

$$\phi_i^G(u_h) = \int_K \varphi_i \vec{a} \cdot \nabla u_h$$

and recast the error as

$$\epsilon(u_h, \psi) = \int_{\Omega_h} \psi_h \vec{a} \cdot \nabla u_h + \sum_{K \in \Omega_h} \sum_{j \in K} \psi_j (\phi_j^K(u_h) - \phi_j^G(u_h))$$

The consistency of the distribution plus the second in (2.6) imply

$$\sum_{j \in K} \phi_j^K = \sum_{j \in K} \phi_j^G = \phi^K$$

If  $C_K$  denotes the number of degrees of freedom of  $K$ , we then rewrite the error as

$$\epsilon(u_h, \psi) = \int_{\Omega_h} \psi_h \vec{a} \cdot \nabla u_h + \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{j, l \in K} (\psi_j - \psi_l) (\phi_j^K(u_h) - \phi_j^G(u_h)) \quad (2.22)$$

The last part consists in using the compactness of  $\psi$  and the continuity of the approximation to write

$$\epsilon(u_h, \psi) = - \int_{\Omega_h} u_h \vec{a} \cdot \nabla \psi_h + \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{j, l \in K} (\psi_j - \psi_l) (\phi_j^K(u_h) - \phi_j^G(u_h)) \quad (2.23)$$

The key is to remember that  $u$  is a classical solution, so that one also has

$$\int_K \varphi_j \vec{a} \cdot \nabla u = 0 \quad \text{and} \quad \int_{\Omega_h} \psi_h \vec{a} \cdot \nabla u = 0$$

and hence

$$\epsilon(u_h, \psi) = - \int_{\Omega_h} (u_h - u) \vec{a} \cdot \nabla \psi_h + \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{j, l \in K} (\psi_j - \psi_l) (\phi_j^K(u_h) - \phi_j^G(u_h - u))$$

So finally

$$|\epsilon(u_h, \psi)| \leq \overbrace{\left| \int_{\Omega_h} (u_h - u) \vec{a} \cdot \nabla \psi_h \right|}^{\text{I}} + \overbrace{\left| \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{j, l \in K} (\psi_j - \psi_l) (\phi_j^K(u_h) - \phi_j^G(u_h - u)) \right|}^{\text{II}}$$

At this point one simply estimates terms. In particular, standard approximation arguments for  $u_h$  and  $\nabla \psi_h$  lead immediately to [63, 105, 11]

$$\text{I} \leq C_1 h^{r+1}$$

Similar arguments allow to deduce

$$|\phi_i^G(u_h - u)| \leq C_2(K) h^{r+2}$$

Lastly, noting that  $\psi_l - \psi_j = \mathcal{O}(h)$ , and that the number of elements in a two-dimensional mesh is of  $\mathcal{O}(h^{-2})$ , one estimates II as

$$\text{II} \leq C(\Omega_h) h^{-2} \times h \times (C_2(K) h^{r+2} + \max_{K,j} |\phi_j^K|) \leq C_4(\Omega_h) h^{r+1} + C_3(\Omega_h) h^{-1} \max_{K,j} |\phi_j^K|$$

So that finally if  $C_0 = \max(C_1, C_4)$ ,

$$|\epsilon(u_h, \psi)| \leq C_0(\Omega_h) h^{r+1} + C(\Omega_h) h^{-1} \max_{K,j} |\phi_j^K|$$

from which the proof follows.  $\square$

A very important building block for high order RD schemes is the following estimate.

**Lemma 2.2.8** (Consistency estimate). *With the same notation of definition 2.2.6, the following estimate holds for the element fluctuation*

$$|\phi^K(u_h)| \leq C(u, \vec{a}, \Omega_h) h^{r+2} \quad (2.24)$$

*Proof.* Since  $u$  is a classical solution :

$$\phi^K(u_h) = \phi^K(u_h - u) = \oint_{\partial K} (u_h - u) \vec{a} \cdot \hat{n}$$

The result follows from standard approximation arguments for  $u_h$  [63, 105, 11].  $\square$

Finally we have :

*Proof of proposition 2.2.5.* Proposition 2.2.5 follows as a corollary of proposition 2.2.7 and of the consistency estimate of Lemma 2.2.8.  $\square$

**Remark 2.2.9** (Approximation and discretization error). *Note that the error (2.20) can be split in two components, as shown by (2.22) and (2.23). The first component is basically the error of the continuous Galerkin scheme, which is the term  $I$ . The magnitude of this term is basically related to the choice of the interpolation. The second component can be interpreted as the additional error introduced by the directional nature of the RD splitting.*

**Remark 2.2.10** (Stability). *The above consistency analysis gives conditions under which that if convergence with respect to the mesh parameter  $h$  is obtained,  $r+1$  convergence rates are obtained w.r.t.  $h$  for a  $ar$ -th degree polynomial approximation, and in correspondence of sufficiently smooth solutions. The missing piece of information is a stability estimate, for example in a finite element sense [105]. For linearity preserving schemes, for example, given two functions  $u$  and  $v$ , while one can easily show a continuity property of the type (cf. definition 2.2.6 for the notation, and [11])*

$$|\beta_i^K \phi^K(u_h) - \beta_i^K \phi^K(v_h)| \leq C \|u_h - v_h\|, \quad \text{with } 0 < C < \infty$$

*we cannot provide a stability statement which ensures e.g. that  $\forall u_h$  in our approximation space*

$$|\sum_K \sum_{j \in K} \beta_j^K u_j \phi^K(u_h)| \geq C' \|u_h\|^2, \quad \text{with } 0 < C' < \infty$$

*If such a stability condition was available, then, using more or less classical arguments [105], we could infer the existence of the discrete solution, and derive more rigorous estimates for the error associated to this solution.*

*Unfortunately, to this day most residual distribution schemes lack a general stability criterion.*

## 2.2.4 Multidimensional upwinding

One of the most original contributions of the work of Roe, Deconinck and collaborators is the construction of *genuinely multidimensional upwind schemes*. To introduce this technique, one must take a closer look at the algebraic form of  $\phi^K$  in the particular case considered. Indeed, on linear triangles, we have on every  $K \in \Omega_h$

$$\nabla \varphi_i|_K = \frac{\vec{n}_i}{2|K|} \tag{2.25}$$

where  $\vec{n}_i$  is the inward pointing normal to the edge in front of node  $i$  (cf. left on figure 2.2), scaled by the length of the edge. As a consequence, the element residual can be recast as

$$\phi^K = \int_K \sum_{j=1}^{j=j_K} \vec{a} \cdot \nabla \varphi_j u_j = \sum_{j=1}^{j=j_K} k_j u_j \quad (2.26)$$

where the *inflow* or *upwind* parameters  $k_j$  are defined as [216, 217, 199, 201]

$$k_j = \frac{\vec{a}_K \cdot \vec{n}_j}{2} = \int_K \vec{a} \cdot \nabla \varphi_j \quad (2.27)$$

with  $\vec{a}_K$  denoting the exact mean value of  $\vec{a}$  over  $K$ . Each  $k_j$  parameter contains the very important information of whether node  $j$  is upstream or downstream w.r.t element  $K$ . In particular, it is evident by the definition of  $\vec{n}_j$  that *if  $k_j$  is positive then node  $j$  is downstream, and if  $k_j$  is negative then node  $j$  is upstream.*

This idea, combined with the underlying linear interpolation, allows an interesting interpretation of what  $\phi^K$  is. In particular, let

$$\begin{aligned} k_j^+ &= \max(0, k_j) = \frac{k_j + |k_j|}{2} \\ k_j^- &= \min(0, k_j) = \frac{k_j - |k_j|}{2} \end{aligned} \quad \text{with} \quad k_j = k_j^+ + k_j^- \quad (2.28)$$

Simple geometry (or equivalently the last in (2.6) and (2.27)) shows that

$$\sum_{j=1}^{j=j_K} k_j = 0 \implies \sum_{j=1}^{j=j_K} k_j^+ = - \sum_{j=1}^{j=j_K} k_j^- \quad (2.29)$$

These identities can be used to prove [202] that *whenever  $u_h$  is linear over  $K$*

$$\begin{aligned} \phi^K &= \int_K \vec{a} \cdot \nabla u_h = \oint_{\partial K} u_h \vec{a} \cdot \hat{n} = N(u_{\text{out}} - u_{\text{in}}) \\ N &= \sum_{j=1}^{j=j_K} k_j^+ = - \sum_{j=1}^{j=j_K} k_j^- \\ u_{\text{out}} &= N^{-1} \sum_{j=1}^{j=j_K} k_j^+ u_j \\ u_{\text{in}} &= -N^{-1} \sum_{j=1}^{j=j_K} k_j^- u_j \end{aligned} \quad (2.30)$$

Note that *provided  $\vec{a}_K \neq 0$ , then  $N > 0$  is well defined.* Using now the linearity of  $u_h$  we can

write now (locally on each  $K \in \Omega_h$ )

$$\begin{aligned}
 u_{\text{out}} &= u_h(\vec{x}_{\text{out}}) \\
 u_{\text{in}} &= u_h(\vec{x}_{\text{in}}) \\
 \vec{x}_{\text{out}} &= N^{-1} \sum_{j=1}^{j=j_K} k_j^+ \vec{x}_j \\
 \vec{x}_{\text{in}} &= -N^{-1} \sum_{j=1}^{j=j_K} k_j^- \vec{x}_j
 \end{aligned} \tag{2.31}$$

The final property easily shown is that

$$\vec{a} \cdot (\vec{x}_{\text{out}} - \vec{x}_{\text{in}}) = |K|N^{-1}\vec{a}_K \cdot \vec{a}_K \geq 0 \tag{2.32}$$

This set of geometrical properties allows to interpret the element fluctuation as *flux difference on the multidimensional streamline crossing the element, and joining the inlet point  $\vec{x}_{\text{in}}$  with the outlet point  $\vec{x}_{\text{out}}$* , and to introduce the following important concept [216, 201]

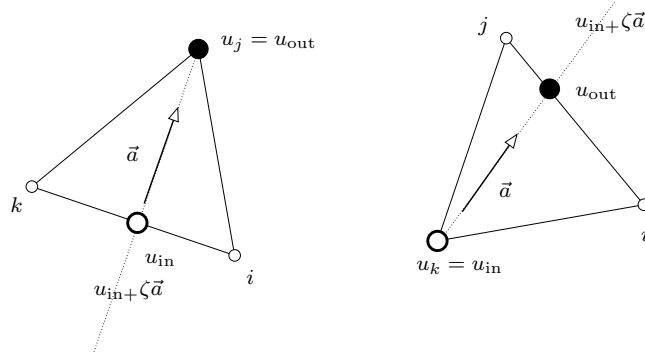


Figure 2.3: Inflow and outflow state. One-target (left) and two-target element (right)

**Definition 2.2.11** (Multidimensional Upwinding). *A Multidimensional Upwind (MU) scheme only distributes to downstream nodes. For a MU scheme*

$$k_i^+ = 0 \Rightarrow \phi_i^K = 0 \tag{2.33}$$

As shown on figure 2.3, the geometrical analysis can be pushed further to show that in two space dimensions two configurations exist [217, 201]<sup>1</sup> :

1. *One target elements* in which the advection speed vector points toward a single node, coinciding with the outflow node. In this case, *all MU schemes distribute the entire residual to this node* ;
2. *Two target elements* in which the advection speed vector points toward an edge facing the only upstream node, coinciding with the inflow node. In this case, *no MU scheme will distribute to this node.*

<sup>1</sup>in three space dimensions there is one more, see [36] for details

One-target elements, in particular, play a special role.

**Remark 2.2.12** (MU schemes and 1-target elements). *MU schemes are locally positive and linearity preserving in one target elements. Positivity is easily checked by considering that, if  $i$  is the only downstream node, then using (2.29)*

$$\phi_i^K = \phi^K = - \sum_{\substack{j \in K \\ j \neq i}} k_j (u_i - u_j)$$

where, by hypothesis  $-k_j \geq 0$  since  $j$  is downstream. Moreover :

$$\beta_i^K = 1, \beta_j^K = 0 \quad \forall j \neq i$$

which is definitely bounded.

## 2.2.5 Scheme zoology : linear schemes

A simple, though important, property of the distribution is the following :

**Definition 2.2.13** (Linear scheme). *A linear scheme is one for which the distribution strategy does not depend on  $u_h|_K$ . In other words, if*

$$\phi_i^K = \sum_{j \in K} c_{ij} u_j$$

the  $c_{ij}$ s do not depend on  $\{u_j\}_{j=1}^{j=j^K}$ .

Unfortunately, the lucky coincidence of remark 2.2.12 does not apply in general [201, 199, 1] :

**Theorem 2.2.14** (Godunov for RD schemes). *Linear RD schemes cannot be simultaneously positive and linearity preserving*

As a consequence of the theorem, linear schemes will be either positive or linearity preserving. In this section we briefly recall some definition of linear splittings. The interested reader can consult [216, 217, 201, 199, 1] for more details.

### Streamline dissipation and Lax-Wendroff schemes

The SUPG scheme of Hughes and co-workers [146, 148] fits in the fluctuation splitting framework. Indeed, neglecting boundary conditions, the SUPG scheme becomes in the  $P^1$  case

$$\begin{aligned} \int_{\Omega_h} \varphi_i \vec{a} \cdot \nabla u_h + \sum_{K \in K_i} \int_K \vec{a} \cdot \nabla \varphi_i \tau \vec{a} \cdot \nabla u_h &= 0 \\ \sum_{K \in K_i} \int_K \varphi_i \vec{a} \cdot \nabla u_h + \sum_{K \in K_i} \int_K \vec{a} \cdot \nabla \varphi_i \tau \vec{a} \cdot \nabla u_h &= 0 \\ \sum_{K \in K_i} \frac{1}{3} \phi^K + \sum_{K \in K_i} \frac{k_i}{|K|} \tau \phi^K &= 0 \end{aligned}$$

Clearly, the SUPG fits in the RD framework with the definition

$$\beta_i^{\text{SUPG}} = \frac{1}{3} + \frac{k_i}{|K|} \tau \tag{2.34}$$

The magnitude of the scaling factor  $\tau$  is such that the *streamline diffusion distribution coefficient*

$$\beta_i^{\text{SD}} = \frac{k_i}{|K|} \tau \quad (2.35)$$

is uniformly bounded [247] (e.g.  $\tau = h_K / \|\vec{a}\|$ , with  $h_K$  the diameter of  $K$ ).

Similarly, the  $P^1$  Taylor-Galerkin scheme obtained by discretizing with a Galerkin scheme the truncated Taylor series expansion [99, 100]

$$\frac{u^{n+1} - u^n}{\Delta t} = \partial_t u^n + \frac{\Delta t}{2} \partial_{tt} u^n = -\vec{a} \cdot \nabla u + \frac{\Delta t}{2} \vec{a} \cdot \nabla (\vec{a} \cdot \nabla u)$$

can be recast as (after lumping of the Galerkin mass matrix)

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \frac{1}{3} \phi^K - \sum_{K \in K_i} \frac{\Delta t k_i}{2|K|} \phi^K \quad (2.36)$$

which fits in the RD framework with definition (2.34) of the distribution coefficients for

$$\tau = \frac{\Delta t}{2}$$

In the RD literature (2.36) is referred to as the Lax-Wendroff (LW) scheme [216, 201, 145].

The streamline upwind discretizations obtained with (2.34) are linear and linearity preserving. The Taylor-Galerkin/LW scheme is second order in space and time.

### First order FV scheme

On the dual mesh composed of the median dual cells, consider the first-order upwind FV scheme for which the discrete counterpart of (2.7)

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{l_{ij} \in \partial C_i} H_{\vec{n}_{ij}}(u_i, u_j)$$

with  $l_{ij}$  is the portion of  $\partial S_i$  separating  $C_i$  from  $C_j$  (see left picture on figure 2.4),  $\vec{n}_{ij}$  is the exterior unit normal to  $\partial S_i$  on  $l_{ij}$ ,  $\vec{n}_{ij} = |l_{ij}| \vec{n}_{ij}$  the scaled exterior normal as in the right picture on figure 2.4, and  $H_{\vec{n}_{ij}}(u, v)$  the upwind numerical flux

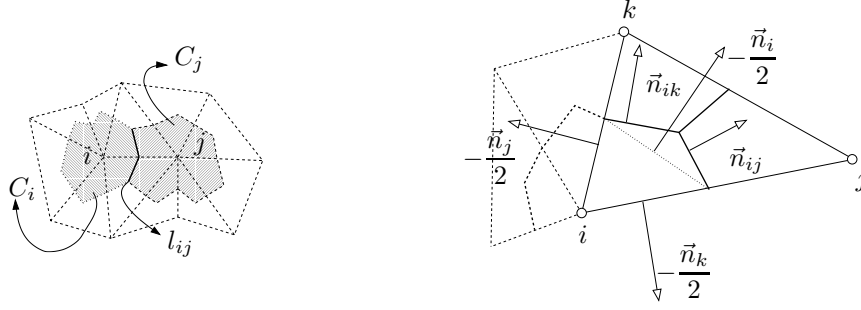
$$H(u, v) = k_{ij} \frac{(u + v)}{2} - \frac{|k_{ij}|}{2} (v - u), \quad k_{ij} = \vec{a} \cdot \vec{n}_{ij} \quad (2.37)$$

The scheme can be easily rewritten in a cell-vertex formalism :

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \sum_{\substack{j \in K \\ j \neq i}} H_{\vec{n}_{ij}}(u_i, u_j)$$

Simple geometrical arguments can be used to recast the right hand side as [201]

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \sum_{\substack{j \in K \\ j \neq i}} (H_{\vec{n}_{ij}}(u_i, u_j) - H_{\vec{n}_{ij}}(u_i, u_i)) \cdot \vec{n}_{ij},$$

Figure 2.4:  $\mathcal{FV}$  scheme. Neighboring cells  $C_i$  and  $C_j$  (left) and cell normals (right)

which is nothing else than the fluctuation splitting scheme

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} = - \sum_{K \in K_i} \phi_i^{\text{FV}}, \quad \phi_i^{\text{FV}} = - \sum_{\substack{j \in K \\ j \neq i}} k_{ij}^-(u_i - u_j) \quad (2.38)$$

where geometry can be used again to show that the  $\phi_j^{\text{FV}}$  verify (2.10).

Definition (2.38) shows that the FV scheme is positive. The reader is referred to [8, 95, 2] for more examples of FV schemes in RD formalism.

### The Lax-Friedrich's scheme

A Lax-Friedrich's splitting is defined by

$$\phi_i^{\text{LF}} = \frac{1}{3} \left( \phi^K + \alpha \sum_{\substack{j \in K \\ j \neq i}} (u_i - u_j) \right) \quad (2.39)$$

The Lax-Friedrich's scheme is positive provided that  $\alpha \geq \max_{j \in K} |k_j|$ .

### The MU N scheme

The N scheme is perhaps the most successful first-order scheme for the solution of the advection equation. First proposed by Roe in the 80's [216], due to its MU character it has the lowest numerical dissipation among first-order schemes [217, 222, 201]. It is defined by the following local nodal residuals:

$$\phi_i^{\text{N}} = k_i^+(u_i - u_{\text{in}}). \quad (2.40)$$

In the 2-target case, a simple geometrical representation exists. Consider the vectors  $\vec{a}_i$  and  $\vec{a}_j$ , parallel to the edges  $\overline{ki}$  and  $\overline{kj}$  respectively, such that  $\vec{a}_i + \vec{a}_j = \vec{a}$  (see figure 2.5). Obviously

$$\phi^K = \phi^K(\vec{a}_i) + \phi^K(\vec{a}_j) = k_i(u_i - u_k) + k_j(u_j - u_k)$$

which immediately gives for the N scheme

$$\phi_i^{\text{N}} = k_i(u_i - u_k) = \phi^K(\vec{a}_i), \quad \phi_j^{\text{N}} = k_j(u_j - u_k) = \phi^K(\vec{a}_j)$$

In the 2-target case, the scheme reduces to first-order upwinding along the edges. The N scheme is positive.



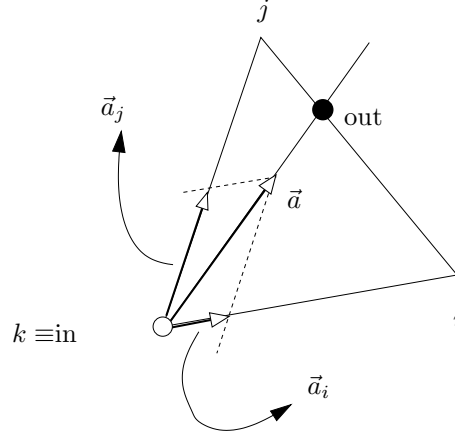


Figure 2.5: Geometry of RD schemes. N scheme in the 2-target case

### The MU LDA scheme

The LDA (Low Diffusion A) is the linear linearity preserving MU scheme defined by the distribution coefficients (cf. equation (2.30)):

$$\beta_i^{\text{LDA}} = k_i^+ N = k_i^+ N^{-1} \quad (2.41)$$

In the 2-target case, a simple geometrical interpretation is possible [201, 216]. As in figure 2.6, we define the sub-triangles  $T_{\text{out}-j-k}$  and  $T_{i-\text{out}-k}$ . Simple trigonometry shows that

$$\begin{aligned} |T_{\text{out}-j-k}| &= \frac{l_{k-\text{out}} k_i}{\|\vec{a}\|} \\ |T_{\text{out}-j-k}| &= \frac{l_{k-\text{out}} k_j}{\|\vec{a}\|} \end{aligned}$$

and that

$$|K| = |T_{\text{out}-j-k}| + |T_{i-\text{out}-k}| = \frac{l_{k-\text{out}}}{\|\vec{a}\|} (k_i + k_j)$$

The distribution coefficients can be written as the area ratios

$$\beta_i^{\text{LDA}} = \frac{k_i}{k_i + k_j} = \frac{|T_{\text{out}-j-k}|}{|K|}, \quad \beta_j^{\text{LDA}} = \frac{k_j}{k_i + k_j} = \frac{|T_{i-\text{out}-k}|}{|K|}$$

Together with the N scheme, the LDA is one of the most successful RD schemes. In particular, the LDA is the RD scheme that has performed better in all the scheme comparisons in published literature [273, 1, 83, 124, 95].

### 2.2.6 Scheme zoology : nonlinear schemes and limiters

As already underlined, linear schemes cannot be positive *and* high order. When approximating solutions containing discontinuities, linear high order schemes introduce dispersive effects otherwise absent in exact solutions of (2.7) [137]. These effects appear in the form

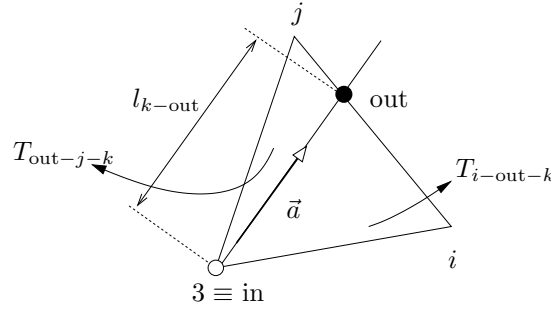


Figure 2.6: Geometry of RD schemes. LDA in the 2-target case

of oscillations in the numerical solution in correspondence of the discontinuity. A nonlinear mechanism has to be introduced to be able to ensure the non-oscillatory character of the numerical solution and go beyond first order of accuracy. We list hereafter the possibilities existing in the RD framework.

### Artificial viscosity

One possibility to reduce the numerical oscillations in correspondence of discontinuities is to introduce a *properly scaled* artificial viscosity term. This is standard practice in the context of SUPG approximations of (2.7) (and of (2.1)) [149, 21, 247]. The main idea is to modify the local splitting as

$$\phi_i^K = \beta_i^K \phi^K + \int_K \nu(u_h) \nabla u_h \cdot \nabla \varphi_i$$

where the last term is an approximation of a Laplacian, and *secret ingredient* is the definition of the numerical viscosity  $\nu(u_h)$ . Normally, the viscosity is proportional to some norm of  $\vec{a} \cdot \nabla u_h$  so that some form of linearity preservation<sup>2</sup> is maintained [149, 21, 247]. Note that due to (2.6)

$$\sum_{j \in K} \int_K \nu(u_h) \nabla u_h \cdot \nabla \varphi_j = 0$$

so that the scheme can still be interpreted as a splitting of  $\phi^K$ .

The approach is quite flexible, being applicable in conjunction with any linear high order scheme. It has of course the draw back of having to *tune* the artificial viscosity. A reappraisal of this approach has been observed in the context of continuous and discontinuous Galerkin approximations [123, 126, 21].

### Gradient orthogonal advection speeds

In the earliest developments of fluctuation splitting schemes, many nonlinear splittings were constructed by introducing an artificial direction parallel to the orthogonal of the solution gradient

$$\vec{a}_{u^\perp} = \frac{(\nabla u_h)^\perp}{\|\nabla u_h\|} \cdot \vec{a}$$

<sup>2</sup>the relevant property is in fact orthogonality, but a proper description of the approach in these terms would require a variational formalism not required by the rest of the manuscript's content

the orthogonal of a vector being in 2d

$$\vec{v}^\perp = (-v_y, v_x)$$

Since  $\vec{a}_{u^\perp} \cdot \nabla u_h = 0$ , (2.7) can be replaced by

$$(\vec{a} + \gamma \vec{a}_{u^\perp}) \cdot \nabla u = 0 \quad (2.42)$$

The key of the recipe here is the local definition of the  $\gamma$  parameter in (2.42). The most interesting formulations are known as the NN scheme [240, 201] and the Level scheme of Roe [217, 201]. In both cases the resulting scheme is multidimensional upwind, thus recovering the properties of remark 2.2.12. In the two target case, the NN scheme is obtained by setting  $\gamma = -1$ , thus only retaining

$$\vec{a}_u = \vec{a} - \vec{a}_{u^\perp}$$

the component of  $\vec{a}$  parallel to  $\nabla u_h$ , and applying the N scheme. Note that for an exact solution  $\vec{a}_u = 0$  so that linearity preservation is maintained.

In the two target case, the Level scheme of Roe attempts at finding a positive splitting that optimizes the time step restriction of proposition 2.2.3, while guaranteeing the boundedness of  $\phi_i^K / \phi^K$ . The construction is slightly more involved, and we refer to [217, 201] for the details.

### PSI schemes and the use of limiters

One of the most successful and original contributions of the fluctuation splitting community is the multidimensional generalization of limited high order schemes.

The idea of applying a limiter function to some ratio of solution variations/local residuals is already embedded in the constructions leading to the NN and Level schemes. However, the first to propose a scheme that was actually obtained via the application of a limiter has been Robert Struijs [239]. Struijs' Positive Streamwise Invariant (PSI) scheme is by far the most successful fluctuation splitting scheme ever. Publications involving its analysis and applications keep on appearing after almost more than 20 years [211, 42, 275, 139, 124, 10, 11, 272, 93, 7, 24, 1, 273, 201]. The PSI scheme is defined by the following formula :

$$\beta_i^{\text{PSI}} = \frac{\max(0, \phi_i^{\text{N}} \phi^K)}{\sum_{j \in K} \max(0, \phi_j^{\text{N}} \phi^K)} \quad (2.43)$$

The PSI scheme reduces to the N scheme in one target cells, while in two target elements it tends to target the nodes that have the largest streamwise variation, that is those for which

$$(u_i - u_{\text{in}}) \phi^K > 0$$

In particular, if  $\phi_i^{\text{N}} \phi^K > 0$

$$\frac{\beta_i^{\text{PSI}} \phi^K}{\phi_i^{\text{N}}} = \frac{(\phi^K)^2}{\phi_i^{\text{N}} \phi^K + \sum_{\substack{j \in K \\ j \neq i}} \max(0, \phi_j^{\text{N}} \phi^K)} = \frac{\phi_i^{\text{N}} \phi^K + \sum_{\substack{j \in K \\ j \neq i}} \phi_j^{\text{N}} \phi^K}{\phi_i^{\text{N}} \phi^K + \sum_{\substack{j \in K \\ j \neq i}} \max(0, \phi_j^{\text{N}} \phi^K)} \leq 1$$

hence

$$\beta_i^{\text{PSI}} \phi^K = \alpha_i \phi_i^{\text{N}}, \quad \alpha_i \in [0, 1]$$

guaranteeing that the sign of coefficients is the same of those of the N scheme. Boundedness is ensured by the renormalization at the denominator, so that the scheme is formally both positive and linearity preserving.

In the two target case, formula (2.43), can be generalized, as realized by Sidilkover and Roe [237]. The idea of the reference is to analyze the two targets case and, given the N scheme split residuals for the two downstream nodes  $i$  and  $j$ , add to these the maximum possible amount of information so that a single target splitting (e.g.  $\beta_i^K = 1, \beta_j^K = \beta_k^K = 0$ ) is obtained as often as possible, while preserving positivity. This is achieved in [237] by setting :

$$\begin{cases} \phi_i^{\text{L}} = \phi_i^{\text{N}} - \phi_j^{\text{N}} \psi(r) \\ \phi_j^{\text{L}} = \phi_j^{\text{N}} - \phi_i^{\text{N}} \psi\left(\frac{1}{r}\right), \quad r = -\frac{\phi_j^{\text{N}}}{\phi_i^{\text{N}}} \end{cases}$$

with  $\psi(\cdot)$  a standard symmetric<sup>3</sup> FV limiter, and the super-script L standing for Limited.

Linearity preservation is formally ensured as long as the limiter verifies the condition  $\psi(1) = 1$ , while positivity requires  $0 \leq \psi(r), \psi(r)/r \leq 1$  [237]. Sidilkover and Roe have even shown that, as in FV schemes, the last condition can be relaxed somewhat on structured grids to  $0 \leq \psi(r), \psi(r)/r \leq 2$ . This analysis opened the door to construction of multidimensional limited nonlinear schemes based on the application of known FV limiters such as Van Albada, SuperBee, and MinMod to the N scheme. In particular, the PSI scheme of Struijs is recovered when employing MinMod [237, 201] !

The application of the same technique to other, simpler, first order positive schemes has been tried since the PhD of Paillere [201, 199] and continued more recently [95, 2]. When using a non MU positive schemes as a basis for the construction, the analysis becomes more complex and the only solution practically working is to set [2]

$$\beta_i^K = \frac{\psi(\beta_i^{\text{P}})}{\sum_{j \in K} \psi(\beta_j^{\text{P}})} \quad (2.44)$$

where positivity is ensured as long as the limiter used verifies  $0 \leq \psi(r) \leq 1$ , and  $\psi(r)/r \geq 0$ . In particular, as in the case of the PSI scheme, (2.44) leads to a discretization verifying

$$\beta_i^K \phi^K = \gamma_i \phi_i^{\text{P}}, \quad \gamma_i \in [0, 1] \quad (2.45)$$

### Blended schemes

A simpler idea has emerged much later : blending a low order scheme with a high order one. Let then  $\beta_i^{\text{HO}}$  be the uniformly bounded distribution coefficient of a linearity preserving RD scheme, and  $\phi_i^{\text{P}}$  the local split residual of a positive linear scheme. A blended scheme reads

$$\phi_i^{\text{B}} = l(u_h) \phi_i^{\text{P}} + (1 - l(u_h)) \beta_i^{\text{HO}} \phi^K \quad (2.46)$$

where not the art is in finding a definition of the *blending parameter*  $l(u_h)$  such that  $l = 1$  in cells containing discontinuities, while  $l = 0$  elsewhere.

<sup>3</sup>symmetry corresponding to the requirement  $\psi(r) = r\psi(1/r)$

The earliest work involving some form of blending [145] is based on a reformulation of the Flux Corrected Transport (FCT) procedure of Boris, Book and Zalesak [39, 38, 40, 282]. More interesting developments have been seen later based on two strategies. One strategy [87, 82] uses the element residual as a regularity monitor taking

$$l(u_h) = \frac{|\phi^K|}{\sum_{j \in K} |\phi_j^P|}$$

While giving acceptable results, this choice is not theoretically sound, positivity not being ensured. Moreover, for steady pure advection it would be more sensible to monitor solution variations in the direction of  $\vec{a}^\perp$ , which are those more responsible for oscillations in steady advection. The quantity  $|\phi^K|$  is instead an index of the local variation parallel to  $\vec{a}$ . Nevertheless, this formulation has encountered some success in literature due to its simplicity and effectiveness in problems containing weak discontinuities [135, 136, 229, 228, 124].

The second approach is to write

$$\phi_i^B = (l(u_h) + (1 - l(u_h))r_i) \phi_i^P, \quad r_i = \frac{\beta_i^{\text{HO}} \phi^K}{\phi_i^P}$$

and to define the blending parameter as  $l(u_h) = l(\{r_j\}_{j \in K})$  such that

$$l(u_h) + (1 - l(u_h))r_i \geq 0 \quad \forall i \in K \quad (2.47)$$

so that the sign of the coefficients of  $\phi_i^P$  is preserved. This approach, that has some similarities with Sidilkover's idea, is followed in [1] using as the N scheme as the positive scheme, and the LDA as the high order one. Explicit formulae satisfying (2.47) are discussed in the reference.

## 2.3 Nonlinear conservation laws and systems

### 2.3.1 Conservative linearizations

When passing to the more general case of (2.1), a very important issue is guaranteeing the correct approximation of discontinuous solutions. Correct approximation here means consistent with the Rankine-Hugoniot conditions associated to (2.1) [231]. When this condition is met, a scheme is said to be *conservative*.

The conditions guaranteeing this consistency have been given in [8, 11], even though the most important of them was known before these papers appeared. The important result is the following (2d version).

**Theorem 2.3.1** (Lax-Wendroff theorem for RD). *Let  $u_0$  be a bounded function  $u_0 \in L^\infty(\mathbb{R}^2)^p$  ( $p$  being the number of equations). Denote by  $\mathcal{C}_h$  the set of dual volumes associated to the degrees of freedom, and let*

$$X_h = \left\{ v_h; v_h|_{C_j} \text{ constant} \in \mathbb{R}^p, \forall C_j \in \mathcal{C}_h \right\}$$

*Let  $u_h$  be the approximation obtained with a RD scheme that satisfies the hypotheses :*

**Hypothesis 1 (Continuity)** *on a triangulation verifying*

$$0 < C_1 \leq \sup_{K \in \Omega_h} \frac{h^2}{|K|} \leq C_2$$

for any  $C \in \mathbb{R}^+$  there exists a  $C'(C, \Omega_h) \in \mathbb{R}^+$  such that for any  $u \in (X_h)^2$  with  $\|u\|_{L^\infty(\mathbb{R}^2)} \leq C$  we have  $\forall K \in \Omega_h$  and  $\forall j \in K$

$$|\phi_j^K| \leq C' h \sum_{l \in K} |u_l - u_j|$$

**Hypothesis 2 (Consistency)** *There exists an approximation  $\mathcal{F}_h$  of the flux such that*

- (i)  $\forall u_h \in X_h$ ,  $\phi^K := \int_K \nabla \cdot \mathcal{F}_h(u_h) = \sum_{j \in K} \phi_j^K$
- (ii)  $\forall u_h \in X_h$ ,  $\forall K_1, K_2$  neighbors :

$$\mathcal{F}_h(u_h)|_{K_1} \cdot \vec{n} = \mathcal{F}_h(u_h)|_{K_2} \cdot \vec{n} \quad \text{a.e. on } K_1 \cap K_2$$

where  $\vec{n}$  is the normal to  $K_1 \cap K_2$

- (iii) For any  $C > 0$ , there exists a  $C'(C)$  such that for any  $u_h \in X_h$  with  $\|u_h\|_{L^\infty(\mathbb{R}^2)} \leq C$  one has for  $K \in \Omega_h$  and  $\mathcal{F}_h^K = \mathcal{F}_h|_K$

$$\|\nabla \cdot \mathcal{F}_h^K\| \leq \frac{C'}{h} \sum_{j,l \in K} |u_j - u_l| \quad \text{a.e. on } K$$

- (iv) For any sequence  $(u_h)_h$  bounded in  $L^\infty(\mathbb{R}^2 \times \mathbb{R}^+)^p$  independently of  $h$  and convergent in  $L^2_{loc}(\mathbb{R}^2 \times \mathbb{R}^+)^p$  to  $u$ , we have

$$\lim_h \|\mathcal{F}_h(u_h) - \mathcal{F}(u)\|_{L^1_{loc}(\mathbb{R}^2 \times \mathbb{R}^+)} = 0$$

Under these hypotheses, and assuming that there exists a constant  $C$  that only depends on  $C_1, C_2$ , and  $u_0$ , and a function  $u \in (L^2(\mathcal{R}^2 \times \mathcal{R}^+))^p$  such that

$$\sup_h \sum_{x,y,t} |u_h(x,y,t)| \leq C$$

$$\lim_h \|u - u_h\|_{L^2_{loc}(\mathcal{R}^2 \times \mathcal{R}^+)^p} = 0$$

Then  $u$  is a weak solution of (2.1).

The proof of this result is reported in [8, 11]. Aside from all technical details, the important point really is hypotheses 2-(i) and 2-(ii). In particular, the two of them together imply that for a RD scheme to be *conservative* then

$$\sum_{j \in K} \phi_j^K = \phi^K = \oint_{\partial K} \mathcal{F}_h(u_h) \cdot \vec{n} \quad (2.48)$$

Because of the upwind nature of the schemes, constantly needing access to the quasi-linear form of the equations, this has been always interpreted in the RD context as a constraint on the linearization to be used in (cf. equation (2.2))

$$\int_K \nabla \cdot \mathcal{F}(u_h) = \int_K \vec{a}(u_h) \cdot \nabla u_h = |K| \vec{a}_K \cdot \nabla u_h|_K, \quad \vec{a}_K = \frac{1}{|K|} \int_K \vec{a}(u_h) \quad (2.49)$$

using the fact that the solution gradient is constant on linear triangles. All the developments of RD schemes for nonlinear problems have relied for many years on the possibility of finding for a given flux function  $\mathcal{F}(u)$  a parameter allowing a simple definition of the exact mean value Jacobian (2.49).

For the Euler equations for a perfect gas, this magic combination is obtained by setting  $\mathcal{F}_h(u_h) = \mathcal{F}(z_h)$  with  $z$  the standard Roe parameter [215]. Because in this case the fluxes are quadratic forms of  $z$ , the element residual can be easily evaluated since

$$\int_K \nabla \cdot \mathcal{F}(\mathbf{z}_h) = \int_K \tilde{\mathbf{a}}_{\mathbf{z}}(\mathbf{z}_h) \cdot \nabla \mathbf{z}_h = |K| \tilde{\mathbf{a}}_{\mathbf{z}}(\mathbf{z}_K) \cdot \nabla \mathbf{z}_h|_K$$

is exact by taking  $\mathbf{z}_K$  as the simple arithmetic average of the nodal values of  $\mathbf{z}$  in  $K$  [86]. Unfortunately, the same is not true anymore on other elements, preventing the simple extension of the method to quadrilaterals and high orders in the nonlinear case.

An interesting approach to go beyond this limitation has been proposed by Abgrall and Barth [6]. The idea proposed in the reference is to keep the definition  $\mathcal{F}_h(u_h) = \mathcal{F}(u_h)$  and introduce a Gaussian quadrature of the quasi-linear form

$$\phi^K = \sum_q \omega_q \tilde{\mathbf{a}}(u_h(x_q)) \cdot \nabla u_h(x_q) \approx \int_K \tilde{\mathbf{a}}(u_h) \cdot \nabla u_h = \oint_{\partial K} \mathcal{F}(u_h) \cdot \vec{n} \quad (2.50)$$

The idea of the reference is that, if the quadrature error is below the truncation error of the scheme, then discrete conservation is practically guaranteed. This intuition is sealed by sound theoretical arguments, including an adaptation of the Lax-Wendroff theorem, and plenty of numerical tests.

The problem of this approach is the following. Every quadrature point defines a quasi-linear form on which a MU scheme can be applied. The linear combination of all of these schemes via the quadrature weights leads to the final conservative discretization. This means that if NQ is the number of quadrature points, the scheme will be NQ times more expensive than normal. In [6] it is shown that with a first order scheme, the number of quadrature points needed to capture correctly a Mach 3.5 shock (still with a simple perfect gas equation of state) is between 7 and 16. This means that, even with an adaptive quadrature strategy, this approach will work in practice, but its use will be impractical for real applications involving high speed flows.

### 2.3.2 Wave decompositions and matrix distribution

The application of the RD approach to a system, requires the definition of the splitting when the residual  $\phi^K$  is a vector.

One approach is to look for some *approximate diagonalization* of the equations. One first introduces a similarity transformation so that locally on each  $K$  one has

$$\tilde{\mathbf{a}}_K \cdot \nabla u_h = L \tilde{\mathbf{a}}_K L^{-1} \cdot \nabla w_h, \quad \nabla w_h = L \nabla u_h$$

The idea is then to look for matrices  $L$  that minimize the off-diagonal entries of  $L \tilde{\mathbf{a}}_K L^{-1}$ , thus maximizing the decoupling, so that scalar schemes can be used on each decoupled problem. This approach tries to find multidimensional characteristic decompositions for the

steady equations [221, 203, 202, 36, 204, 178, 85], and it bears many common points with preconditioning and characteristic time stepping techniques [260].

For the Euler equations, this technique has led to the powerful *hyperbolic-elliptic splitting* and *potential decomposition* [193, 194, 209, 208, 202, 36] approaches, allowing to completely decouple the equations in the supersonic case, and to treat separately the hyperbolic and elliptic sub-components in the subsonic flows. This approach yields a very fast discretization, allowing to recover correctly the low Mach limit (and even potential solutions), and guaranteeing the monotonicity of the results in presence of discontinuities.

While yielding results very competitive w.r.t the best finite volume schemes [229, 124], the drawback of the approximate diagonalizations is its lack of generality. A different approach, proposed in [256, 257], is to formally generalize the linear distribution schemes, by introducing a matrix notation. To give an example, for the Euler equations for a perfect gas, one has

$$\phi^K = \sum_{j \in K} k_j \mathbf{z}_j, \quad k_j = \frac{1}{2} \vec{a}_{\mathbf{z}}(\mathbf{z}_K) \cdot \vec{n}_j$$

where  $\mathbf{z}$  is the Roe parameter and now  $\vec{a}_{\mathbf{z}}(\mathbf{z}_K) \cdot \vec{n}_j$  is a matrix (cf equation (2.2)). The hyperbolic character of the equations allows to define the  $k_j^\pm$  parameters via matrix decomposition, and apply any of the schemes of section §2.2.4. Preconditioning techniques can still be applied to improve the iterative convergence [260, 257, 255, 52, 53, 136]. This approach has encountered an important success and is the one used in the work reported here.

## 2.4 Time dependent problems

When using (2.11) to compute time dependent solutions, first order of accuracy is observed in space, and improving the accuracy of the time integration scheme does not improve the convergence rates observed [239, 145].

This fact has been always known to be related to some form of weak inconsistency. The first explanation has been proposed by the group at the von Karman Institute [174]. The idea is to rewrite RD schemes as a Petrov-Galerkin (PG) finite element scheme obtained by perturbing the continuous Galerkin discretization with a locally constant term :

$$\beta_i^K \phi^K = \int_K \varphi_i \vec{a} \cdot \nabla u_h + \int_K \alpha_K^i \vec{a} \cdot \nabla u_h \quad (2.51)$$

In the  $P^1$  case<sup>4</sup>, one readily finds out that if  $\alpha_K^i = \beta_i^K - 1/3$ , then (2.51) is satisfied. When going to time dependent problems, the PG analogy leads to the appearance of a mass matrix :

$$\begin{aligned} \int_{\Omega_h} \varphi_i (\partial_t u_h + \vec{a} \cdot \nabla u_h) + \sum_{K \in K_i} \int_K \alpha_K^i (\partial_t u_h + \vec{a} \cdot \nabla u_h) &= 0 \\ \Rightarrow \sum_{K \in K_i} \sum_{j \in K} m_{ij}^K \frac{du_j}{dt} + \beta_i^K \phi^K &= 0 \end{aligned} \quad (2.52)$$

---

<sup>4</sup>and if  $\vec{a}$  is constant, or if the locally linearized problem is analyzed...



where the PG mass matrix is

$$m_{ij}^K = \int_K (\varphi_i + \alpha_K^i) \varphi_j = \frac{|K|}{3} \left( \beta_i^K + \frac{3\delta_{ij} - 1}{12} \right)$$

The original update scheme (2.11) can be now obtained by lumping the Petrov-Galerkin mass matrix. The loss of accuracy is then attributed to mass lumping, as largely confirmed by numerical experiments [174]. The construction has been extended to the Euler equations in [107].

Roughly at the same time, at the Lund University in Sweden, Doru Caraeni argued that the missing ingredient in (2.11) is the linearity preservation property. In particular, he proposed to redefine the element residual in the time dependent case as [54]

$$\Phi^K = \int_K \left( \partial_t u_h + \vec{a} \cdot \nabla u_h \right) = \sum_{j \in K} \frac{|K|}{3} \frac{du_j}{dt} + \phi^K$$

He shows that schemes defined by

$$\sum_{K \in K_i} \beta_i^K \Phi^K = 0 \tag{2.53}$$

yield indeed high order results for time dependent problems. Schemes (2.53) and (2.52) however *are not* the same.

These two experiences, however, suggest that :

- the recipe for reaching high order is to maintain some sort of *orthogonality* property : if the discrete solution is replaced by an exact classical solution, the discrete equations reduce to an identity ;
- genuinely explicit schemes cannot exist since whatever bias is present in splitting has to be applied to the time derivative as well, requiring the inversion of a mass matrix.

The second point has led to some developments based on the explicit Lax-Wendroff scheme (2.36) of section §2.2.4 [145]. The author of this manuscript has himself taken some part to these developments [RD99, CdSR<sup>+</sup>00] whose objective was to look for some form of non-linear correction to apply the LW scheme to problems containing discontinuities. However, as already mentioned, (2.36) is a Galerkin discretization, and these developments did not actually bring further understanding of why both (or any of) (2.52) and (2.53) work.

## 2.5 Beyond second order of accuracy

As the truncation error analysis of section §2.2.2 shows, to go beyond second order of accuracy one should increase the degree of the polynomial approximation. The first published examples of more than second order schemes appeared in [11] and [54]. The two developments are based on different ideas.

In [11], the authors make use of higher degree  $P^k$  Lagrange triangles. In every triangle, a  $P^1$  conformal sub-triangulation is introduced (cf. figure 2.7), and sub-elemental fluctuations, computed using the higher degree approximation, are distributed to three nodes of the sub-element<sup>5</sup>. Linearity preserving schemes such as the LDA are easily extended to this

<sup>5</sup>a very similar idea was also proposed in [84]

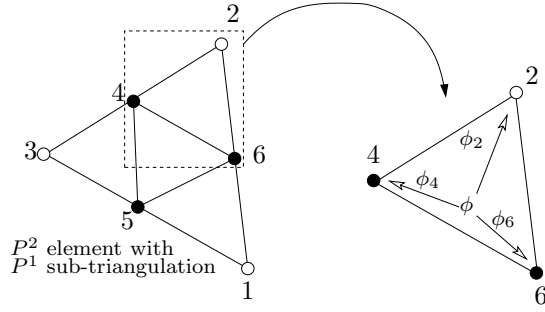


Figure 2.7: P2 element (left) and sub-element distribution (right)

framework. A PSI scheme is proposed that is constructed by

1. define a  $P^1$  sub-elemental N scheme. In the example on figure 2.7 (cf. equation (2.40)) :

$$\phi_2^N = (k_2^{246})^+ (u_2 - u_{in}^{246})$$

2. apply formula (2.43) to the three N scheme split residuals replacing  $\phi^K$  by the sub-elemental higher order fluctuation. For example, on the right in figure 2.7

$$\beta_2^{246} = \frac{\max(0, \phi_2^N \phi^{246})}{\sum_{j \in \{2,4,6\}} \max(0, \phi_j^N \phi^{246})}, \quad \phi^{246} = \int_{246} \vec{a} \cdot \nabla u_h^{P^2}$$

The approach proposed by Doru Caraeni is instead to reconstruct nodal gradients, and, using these, compute the fluctuation as

$$\phi^K = \oint_{\partial K} u_h \vec{a} \cdot \vec{n}$$

where the knowledge of the nodal gradients allows to increase the accuracy with which to edge integrals in the last formula are evaluated. The simple reconstruction formula

$$\nabla u_i = \frac{\sum_{K \in K_i} |K| \nabla u_h^{P^1} \Big|_K}{\sum_{K \in K_i} |K|}$$

can be used to achieve third order of accuracy [54, 193]. A nonlinear blended scheme based on this approach has been proposed in [136].



## Chapter 3

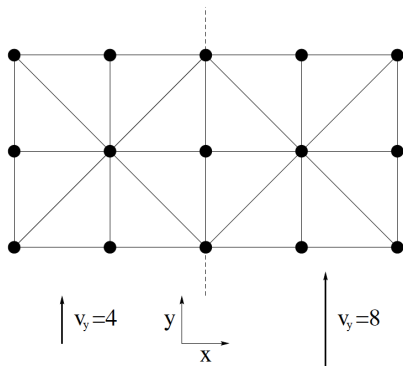
# Contributions : conservative higher order schemes for steady state problems

### 3.1 Conservation via direct flux approximation

The need of a local exact mean value linearization of the flux Jacobian to retain discrete conservation represented a major limitation for fluctuation splitting. Discrete conservation could not be achieved in many situations of interest :

1. in the application of RD schemes to other systems such as the MHD equations [82] or even Shallow Water flows [141] :
2. in the the application of RD schemes to non-triangular elements [6]

The use of Roe's linearization brings other undesired effects. These are related to the fact that discrete gradients of physical quantities are derived from the gradients of the components of Roe's parameter :



$$z = \sqrt{\rho} \begin{bmatrix} 1 \\ \vec{v} \\ H \end{bmatrix}$$

$\vec{v}$  being the flow speed, and  $H$  the total enthalpy

$$H = \frac{\gamma}{\gamma - 1} \frac{p}{\rho} + \frac{\vec{v} \cdot \vec{v}}{2}$$

with  $p$  the pressure and  $\rho$  the density. The linear approximation of  $z$  carries some unphysical effects in simple situations such as the supersonic steady mesh aligned contact depicted on the left with constant pressure and density, but discontinuous vertical speed.

The application of the standard blended scheme that uses Roe's linearization to this configuration leads to the results reported on figure 3.1 : spurious variations are observed

which *increase if the mesh is refined*. A similar behavior is obtained with the N scheme, while the simulation with the LDA scheme blows up (negative pressure). This is quite a disappointing feature.

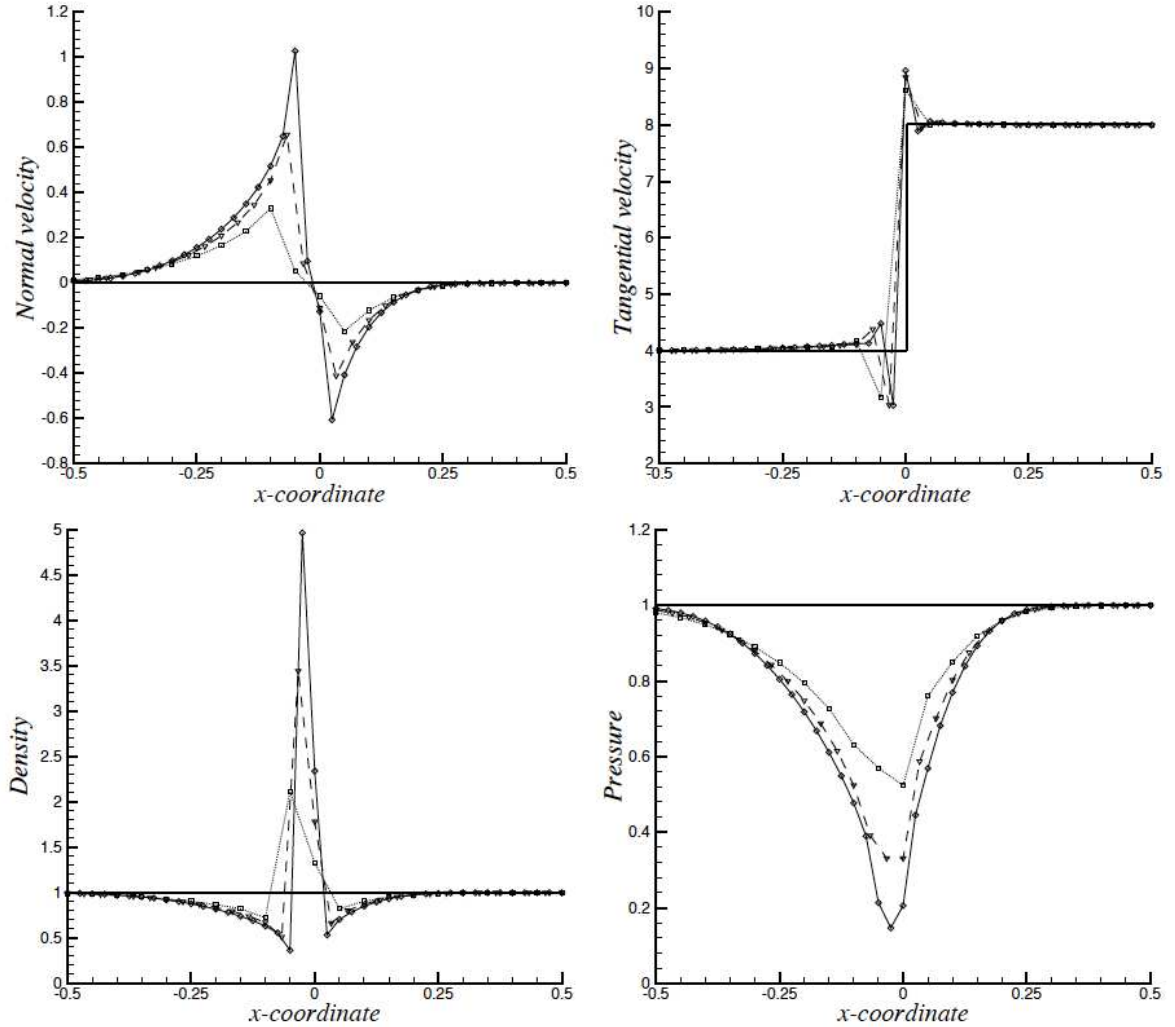


Figure 3.1: Constant density/pressure contact : outlet solution of the N scheme and Roe's linearization on  $40 \times 40$  (squares),  $60 \times 60$  (triangles), and  $80 \times 80$  (diamonds) meshes [CRD02]

A solution of this problem is suggested by hypotheses 2-(i) and 2-(ii) and by (2.48) : discrete conservation is retained is a continuous flux approximation  $\mathcal{F}_h$  exists such that

$$\sum_{j \in K} \phi_j^K = \oint_{\partial K} \mathcal{F}_h \cdot \vec{n} \quad (3.1)$$

Theorem 2.3.1 does not require this approximation to be  $\mathcal{F}(u_h)$  or to be anyhow related to how the splitting is performed, provided (3.1) is satisfied.

### 3.1.1 Boundary integration of the flux and flux approximation

In [RCD01, CRD02] (and in the related work [79, QRCD02, RCD04, RCD05, Ric05]) we proposed to compute the element fluctuation  $\phi^K$  by directly integrating the right hand side of (3.1). Even though this was not clear at the time, the quadrature formulae used on the faces of  $K$  are directly linked to a continuous polynomial reconstruction of the flux. This is immediately seen by setting for example

$$\mathcal{F}_h = \sum_{\sigma=1}^{\sigma^{P^k}} \mathcal{F}_\sigma \varphi_\sigma^{P^k} \Rightarrow \nabla \cdot \mathcal{F}_h|_K = \sum_{\sigma=1}^{\sigma^{P^k}} \mathcal{F}_\sigma \cdot \nabla \varphi_\sigma^{P^k} \quad (3.2)$$

where the flux values are reconstructed using the underlying solution representation :

$$\mathcal{F}_\sigma = \mathcal{F} \left( \sum_{j \in K} \varphi_j(\vec{x}_\sigma) u_j \right)$$

We consider the following cases :

**$P^1$  flux** In this case there are three flux values to evaluate, coinciding with the values at the elements vertices, and the  $\varphi_\sigma^{P^1}$  are linear, hence  $\nabla \cdot \mathcal{F}_h|_K$  is constant and given by (cf. equation (2.6) and figure 2.2)

$$\nabla \cdot \mathcal{F}_h|_K = \sum_{\sigma=1}^3 \mathcal{F}_\sigma \cdot \nabla \varphi_\sigma^{P^k} = \sum_{j \in K} \frac{1}{2|K|} \mathcal{F}_j \cdot \vec{n}_j$$

The geometrical relation  $\sum_{j \in K} \vec{n}_j = 0$  leads to the result that

$$\int_K \nabla \cdot \mathcal{F}_h = \sum_{f \in \partial K} \sum_{p \in f} \frac{\mathcal{F}_p}{2} \cdot \vec{n}_f$$

with  $\vec{n}_f$  the outward normal to the face  $f \in \partial K$ , scaled by its length. Last expression is the result of the integration of the right hand side of (3.1) with trapezium rule.

**$P^2$  flux** In this case there are six flux values, corresponding to the three vertices and the three edge mid-points, the gradients of the flux shape functions  $\varphi_\sigma^{P^2}$  are linear, and so is  $\nabla \cdot \mathcal{F}_h|_K$ . For the shape functions we have

$$\begin{aligned} \nabla \phi_\sigma^{P^2} &= (4\varphi_\sigma^{P^1} - 1) \frac{\vec{n}_\sigma}{2|K|} && \text{for a vertex } \sigma \\ \nabla \phi_{\sigma_m}^{P^2} &= 4\varphi_{\sigma_1}^{P^1} \frac{\vec{n}_{\sigma_2}}{2|K|} + 4\varphi_{\sigma_2}^{P^1} \frac{\vec{n}_{\sigma_1}}{2|K|} && \text{for a midpoint } \sigma_m = (\sigma_1 + \sigma_2)/2 \end{aligned}$$

The integral of each of the  $\varphi_\sigma$ s is equal to  $|K|/3$ . In both cases (after a little algebra) we have

$$\int_K \nabla \cdot \mathcal{F}_h = \sum_{f \in \partial K} \left( \frac{1}{6} \mathcal{F}_{\sigma_1} + \frac{2}{3} \mathcal{F}_{\sigma_m} + \frac{1}{6} \mathcal{F}_{\sigma_2} \right) \cdot \vec{n}_f$$

corresponding to use of Simpson's rule on each face of the element.

A similar analysis shows that the 3/8 Simpson's rule on the faces is equivalent to a  $P^3$  flux. Note that formulas of equivalent accuracy can be of course fit in the same case. For example, the use of the standard 2 point Gauss formula on each edge is also equivalent to the exact integration of a  $P^3$  flux. Similar constructions can be obtained by resorting to Raviart-Thomas approximation of the flux, as recently proposed in the context of Spectral Difference schemes [177, 266].

### 3.1.2 First order and high order schemes

The question that has to be answered is how to perform the splitting. The answer to this question is given by the left hand side of (3.1) : one just needs to guarantee that

$$\sum_{j \in K} \phi_j^K = \phi^K$$

For high order linear schemes it is enough to set

$$\phi_i^K = \beta_i^K \phi^K = \beta_i^K \oint_{\partial K} \mathcal{F}_h \cdot \vec{n}$$

evaluating the  $\beta_i^K$  matrix in whatever averaged state is more convenient. The consistency condition

$$\sum_{j \in K} \beta_j^K = 1$$

respected by all linearity preserving schemes, automatically ensures discrete conservation.

Concerning linear first order schemes, the first order FV scheme can be immediately recast into a conservative form fitting into this framework (see [DR07] and references therein for details), while the Lax-Friedrich's distribution is also easily modified as

$$\phi_i^{\text{LF}} = \frac{1}{3} \left( \oint_{\partial K} \mathcal{F}_h \cdot \vec{n} + \alpha \sum_{\substack{j \in K \\ j \neq i}} (u_i - u_j) \right)$$

The case of real interest is the N scheme, for which a solution is proposed in [CRD02] : it suffices to replace the inflow state  $u_{\text{in}}$  in (2.40) by a conservative state  $u_c$  ensuring conservation. In formulas (cf. equation (2.30) for the notation) :

$$\begin{aligned} \phi_i^{\text{N}} &= k_i^+(u_i - u_c) \quad \text{plus} \quad \sum_{j \in K} k_j^+(u_j - u_c) = \phi^K \\ &\Rightarrow u_c = \overline{u_{\text{out}}} - N^{-1} \phi^K \end{aligned} \tag{3.3}$$

This simple formula allows to extend the use of the multidimensional upwind N scheme to this conservative setting. In particular in (3.3) the flux Jacobians needed to evaluate the  $k_j$  parameters can be computed using any arbitrary linearization.

As an example, figures (3.2) and (3.3) show the comparison between the results obtained with the Roe linearization and the conservative flux based approach for a Mach 10 bow shock flow around a circular cylinder. The shock is computed with the correct strength and position, with no spurious oscillations. The blue results in the figures are computed using arithmetic averages of pressure, density and velocity to evaluate the flux Jacobians.

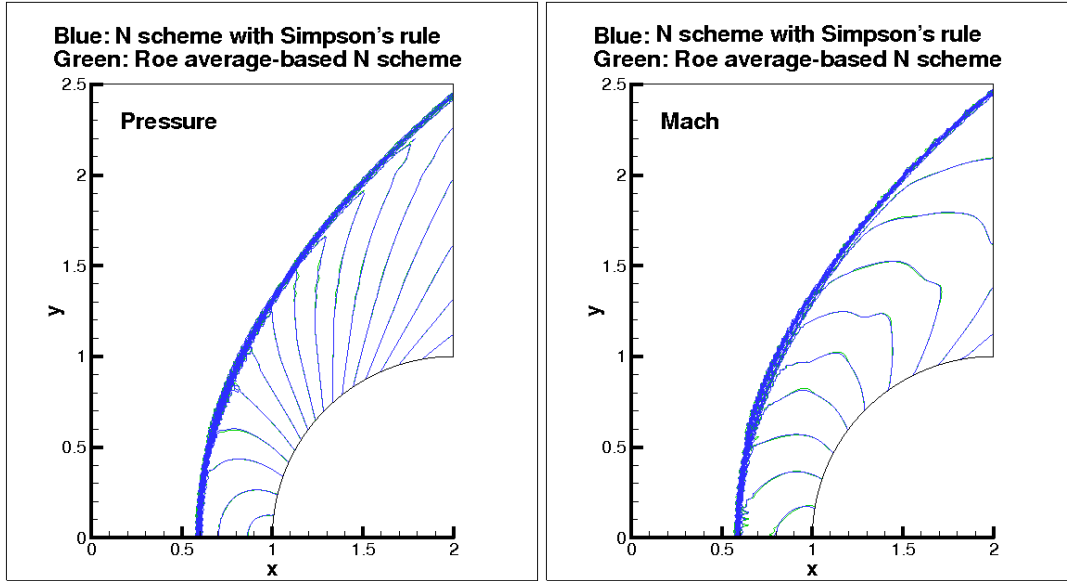


Figure 3.2: Mach 10 bow shock : N scheme. Pressure and Mach number contours

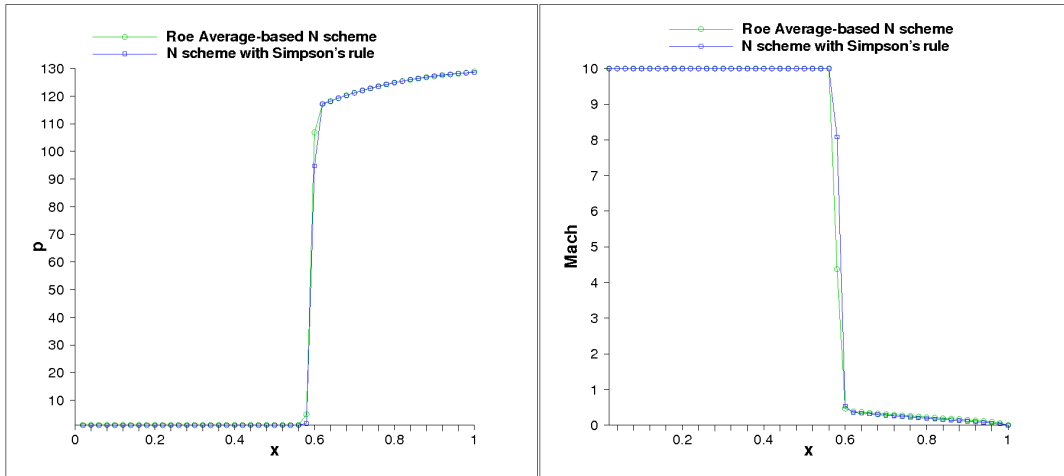


Figure 3.3: Mach 10 bow shock : N scheme. Pressure and Mach along the stagnation line

A nonlinear variant of the conservative N scheme can be obtained using limiter (2.44) . In particular, following [10] let  $\boldsymbol{\ell}_k$  and  $\mathbf{r}_k$  be the  $k$ -th left and right eigenvector of the flux Jacobian  $\bar{\mathbf{a}}(\bar{\mathbf{u}}) \cdot \hat{\mathbf{v}}$ , with  $\bar{\mathbf{u}}$  obtained by simple arithmetic averaging, and  $\hat{\mathbf{v}}$  the local direction of the velocity. We proceed as follows

1. Define scalar residuals ( $^t$  denoting the transpose operator)

$$(\varphi_i^N)_k = \boldsymbol{\ell}_k^t \phi_i^N \quad \text{and} \quad (\varphi_i^K)_k = \boldsymbol{\ell}_k^t \phi_i^K \quad (3.4)$$



2. Set

$$(\varphi_i)_k = \frac{\max(0, (\varphi_i^N)_k (\varphi^K)_k)}{\sum_{j \in K} \max(0, (\varphi_j^N)_k (\varphi^K)_k)} (\varphi^K)_k \quad (3.5)$$

3. Transform to physical variables

$$\phi_i^K = \sum_k \mathbf{r}_k (\varphi_i)_k \quad (3.6)$$

The result obtained with this limited N scheme (LN scheme) for the Mach 10 bow shock is reported on figure 3.4 (see [DR07, RCD05, Ric05] for more details and results).

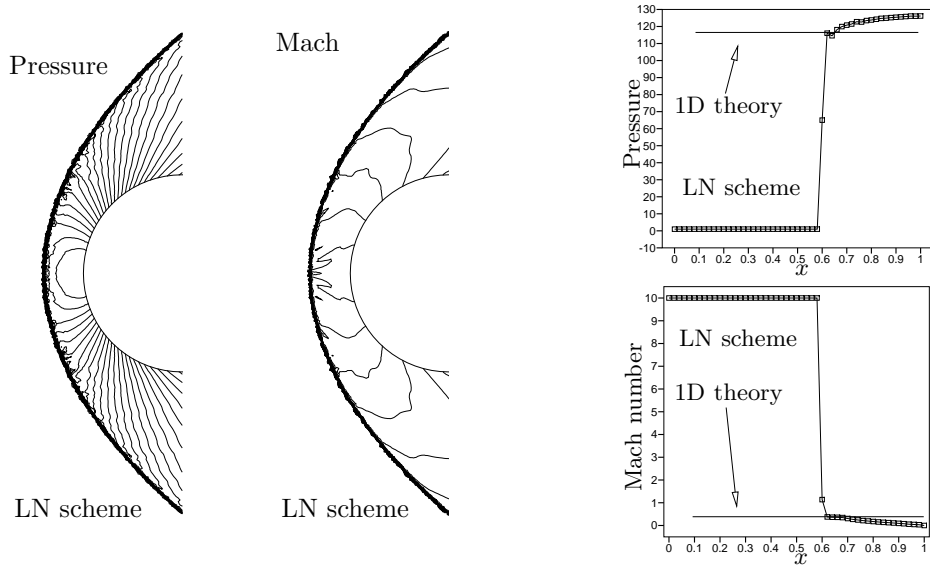


Figure 3.4: Mach 10 bow shock around a circular cylinder : LN scheme. Pressure (left) and Mach (right) contours, and pressure and Mach distribution along the stagnation line

### 3.1.3 Additional observations

The first interesting result is the following.

**Proposition 3.1.1** (Flux based schemes and steady contacts). *Mesh aligned steady contact discontinuities are preserved by high order RD based on direct flux approximation, provided  $\vec{v} \cdot \vec{n}|_{\partial K} = 0$ , and provided that direct interpolation of the pressure in the flux is used.*

*Proof.* If the pressure is approximated directly, then for the Euler equations

$$\phi^K = \oint_{\partial K} (u\vec{v})_h \cdot \vec{n} + \oint_{\partial K} \begin{bmatrix} 0 \\ p\vec{n} \\ p\vec{v} \cdot \vec{n} \end{bmatrix}_h = 0$$

since the pressure is constant across the contact, since by hypothesis  $\vec{v} \cdot \vec{n}|_{\partial K} = 0$ , and using the fact that  $\oint_{\partial K} \vec{n} = 0$ . As a consequence, high order RD will preserve this state indefinitely.  $\square$

This property (easily verified numerically) removes on a major flaw of the RD approach.

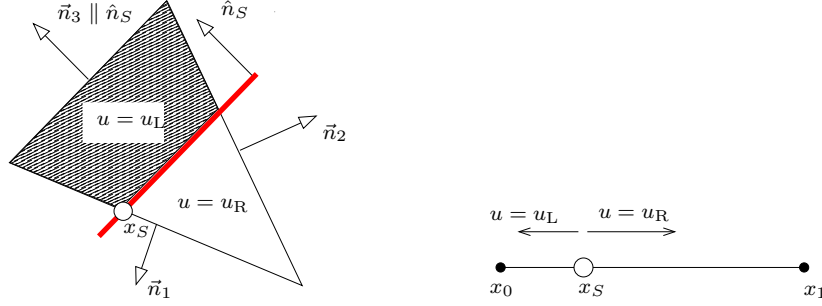


Figure 3.5: Mesh aligned discontinuity

Another property that is easily checked numerically is the following.

**Proposition 3.1.2** (Flux approximation and trapped shocks). *Steady mesh aligned discontinuities are trapped and preserved by high order RD schemes if the polynomial degree of the flux approximation is the same of the solution.*

*Proof.* Consider the situation of figure 3.5 : all the degrees of freedom in the shaded area verify  $u = u_L$ , and those in the white area verify  $u = u_R$ . By definition of the steady discontinuity, however, all of them verify :

$$\mathcal{F}(u_j) \cdot \hat{n}_S = \mathcal{F}_0 = \text{const} \quad \text{and}$$

Conversely, we can set

$$\mathcal{F}(u_L) \cdot \hat{n}_S^\perp = \mathcal{F}_L^\perp \quad \text{and} \quad \mathcal{F}(u_R) \cdot \hat{n}_S^\perp = \mathcal{F}_R^\perp$$

Note also that in this configuration, the shock normal is aligned with  $\vec{n}_3$ , and hence  $\vec{n}_3 \cdot \hat{n}_S^\perp = 0$ . We can write for the element fluctuation :

$$\phi^K = \sum_{j \in K} \left( \int_{f_1} \varphi_j \mathcal{F}_j \cdot \vec{n}_1 + \int_{f_2} \varphi_j \mathcal{F}_j \cdot \vec{n}_2 + \int_{f_3} \varphi_j \mathcal{F}_j \cdot \vec{n}_3 \right)$$

We can now use the fact that

$$\vec{n}_j = \vec{n}_j \cdot \hat{n}_S \hat{n}_S + \vec{n}_j \cdot \hat{n}_S^\perp \hat{n}_S^\perp = n_j^S \hat{n}_S + n_j^{S^\perp} \hat{n}_S^\perp$$

and that, for the type of elements considered here  $\varphi_j = 0$  if  $j \notin f_l$ , to write

$$\phi^K = \sum_{l=1}^3 \sum_{j \in f_l} \int_{f_l} \varphi_j \mathcal{F}_0 n_l^S + \sum_{l=1}^2 \sum_{\substack{j \in f_l \\ u_j = u_L}} \int_{f_l} \varphi_j \mathcal{F}_L^\perp n_l^{S^\perp} + \sum_{l=1}^2 \sum_{\substack{j \in f_l \\ u_j = u_R}} \int_{f_l} \varphi_j \mathcal{F}_R^\perp n_l^{S^\perp}$$

having used the hypothesis  $\vec{n}_3 \cdot \hat{n}_S^\perp = 0$  (cf. figure 3.5). Since  $\sum_l \vec{n}_l = 0$ , the first term vanishes identically :

$$\sum_{l=1}^3 \sum_{j \in f_l} \int_{f_l} \varphi_j \mathcal{F}_0 n_l^S = \mathcal{F}_0 \sum_{l=1}^3 \int_{f_l} n_l^S = 0$$

The last two terms also vanish for symmetry reasons. Indeed, with the notation of the right picture on figure 3.5, these integrals can be recast as

$$\sum_{\substack{j \in f_l \\ u_j = u_R}} \int_{f_l} \varphi_j \mathcal{F}_R^\perp n_l^{S^\perp} = (\mathcal{F}_R^\perp n_l^{S^\perp}) \sum_{\substack{j \in f_l \\ x_j < x_S}} \int_{x_0}^{x_1} \varphi_j = (\mathcal{F}_R^\perp n_l^{S^\perp}) I_l^R$$

and similarly for the  $\mathcal{F}_L$  terms. The line integrals are the same on both faces, the Lagrange shape functions reducing to their one dimensional form on each face. Moreover, the number of degrees of freedom in the pre- and post-discontinuity are also the same on the two faces due to the mesh alignment of the discontinuity. As a consequence  $I_1^R = I_2^R = I^R$ , and  $I_1^L = I_2^L = I^L$ . This fact, plus the identity  $n_1^{S^\perp} = -n_2^{S^\perp}$  leads to

$$\sum_{l=1}^2 \sum_{\substack{j \in f_l \\ u_j = u_L}} \int_{f_l} \varphi_j \mathcal{F}_L^\perp n_l^{S^\perp} + \sum_{l=1}^2 \sum_{\substack{j \in f_l \\ u_j = u_R}} \int_{f_l} \varphi_j \mathcal{F}_R^\perp n_l^{S^\perp} = \mathcal{F}_L^\perp I^L (n_1^{S^\perp} + n_2^{S^\perp}) + \mathcal{F}_R^\perp I^R (n_1^{S^\perp} + n_2^{S^\perp}) = 0$$

hence  $\phi^K = 0$ . If the mesh is aligned with the discontinuity, then this is true  $\forall K$ , and the discontinuity is trapped.  $\square$

**Remark 3.1.3** (Choice of the approximation). *Even if the last result might seem a positive one, it actually is not. The practice shows that expansion shocks are as easily trapped as compression ones, independently on the splitting strategy.*

*Face quadrature formulas consistent with polynomials of at least one degree higher than the solution should be used in practice. This allows to break these unphysical shocks, even though not completely forbidding their appearance under other forms, as shown in [230].*

A last interesting observation is that when replacing back  $u_c$  in the N scheme splitting (3.3) one obtains with few algebraic manipulations (cf. equation (2.41))

$$\phi_i^N = \beta_i^{\text{LDA}} \phi^K + d_i^N \quad (3.7)$$

with

$$d_i^N = k_i^+(u_i - u_{\text{out}}) = \sum_{\substack{j \in K \\ j \neq i}} k_i^+ N^{-1} k_j^+(u_i - u_j) \quad (3.8)$$

**Remark 3.1.4** (Relations between N and LDA). *The term  $d_i^N$  is a dissipation term. The last relation shows that the N scheme is obtained by adding to the high order LDA scheme a cross-wind dissipation term.*

*Proof.* Indeed, one easily checks in the scalar case that

$$\sum_{j \in K} u_j d_j^N = \epsilon^N = \frac{1}{2} \sum_{j, l \in K} (u_j - u_l) k_j^+ N^{-1} k_l^+(u_j - u_l) \geq 0$$

A similar result is obtained for a symmetric system by analyzing the bi-linear form associated to the block matrix (see [DR07, 20, 6] for details)

$$D^N = \begin{bmatrix} k_1^+ & 0 & 0 \\ 0 & k_2^+ & 0 \\ 0 & 0 & k_3^+ \end{bmatrix} - \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix} N^{-1} \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix}^T \quad (3.9)$$

The fact that it is a cross-wind dissipation is easily seen from the fact that it only involves differences between downstream nodes.  $\square$

The last remark suggests a similar description for the LF scheme. Indeed we have

$$\begin{aligned} \phi_i^{\text{LF}} &= \frac{\alpha}{3}(u_i - u_c) \quad \text{plus} \quad \sum_{j \in K} \frac{\alpha}{3}(u_j - u_c) = \phi^K \\ \Rightarrow u_c &= \bar{u} - \frac{1}{\alpha}\phi^K, \quad \bar{u} = \frac{1}{3} \sum_{j \in K} u_j \end{aligned} \quad (3.10)$$

which is equivalent to (2.39).

## 3.2 Higher order schemes for steady conservation laws

### 3.2.1 Generalities

Several numerical discretizations have shown potential for increasing the accuracy way beyond second order : stabilized continuous Galerkin and Petrov-Galerkin schemes [21, 271, 198, 60, 280, 50, 49], discontinuous Galerkin (DG) schemes [69, 69, 68, 66, 71] (and see the review [234] and references therein), schemes based on WENO reconstructions (see [233] and references therein), the spectral finite volume schemes of [263, 264, 265, 267], the residual based compact schemes of [167, 168, 75, 76, 73, 74, 77], and residual distribution [11, 54].

The main motivation for seeking such an increase of accuracy is related to efficiency : higher order methods make better use of the discrete unknowns. In other words, a  $k$ -th order method is more efficient in terms of the ratio

$$\eta_k = \frac{1}{\text{error} \times \text{CPU time}} = \frac{1}{c_{\text{err}} h^k \times c_{\text{CPU}} n_{\text{op.s}}} = \eta_{\text{scheme}}^k \frac{1}{h^k n_{\text{dof}}^k} \quad (3.11)$$

assuming that the time necessary for one operation  $c_{\text{CPU}}$  is universal<sup>1</sup>. The factor  $\eta_{\text{scheme}}$  is the inverse of a cost to obtain a unit error with one degree of freedom. The bigger  $\eta_{\text{scheme}}$ , the smaller this unit cost, the better the scheme.

The efficiency factor  $\eta_k$  is a telltale of the runtime a method needs to achieve a fixed error level : the higher  $\eta_k$  the better the method.

This suggests that increasing  $k$  is beneficial for reducing the time necessary to achieve a certain error. A verification of (3.11) for the DG method is presented in [65]. The comparison of second and third order schemes shows that the gains in efficiency when increasing the accuracy can be of several orders of magnitude. Indeed, the relation

$$\frac{\eta_{k+m}}{\eta_k} = \frac{\eta_{\text{scheme}}^{k+m}}{\eta_{\text{scheme}}^k} \frac{n_{\text{dof}}^k}{n_{\text{dof}}^{k+m}} h^{-m} \quad (3.12)$$

<sup>1</sup>which is not, but this is not part of the topics covered in the manuscript

shows that, at least on fine meshes, the reduction in accuracy obtained for  $m \geq 1$  will overwhelm the additional cost of the increase in accuracy related possibly to the larger number of degrees of freedom used, and certainly to the higher unit cost  $1/\eta_{\text{scheme}}^{k+m}$ .

These arguments, and the promising results of the schemes mentioned in the beginning of this section, has motivated more and more researchers to look into higher order schemes. Under the increasing interest of the aeronautic industry, a certain number funding actions have accelerated these developments. One of the most ambitious of these projects is the EU STREP project ADIGMA, funded in the framework of the call Call: FP6-2005-Aero-1<sup>2</sup>. The objective of ADIGMA, as stated on the EU page, was “to add a major step towards the development of next-generation CFD tools for advanced aerodynamic applications with significant improvements in accuracy and efficiency.” In this context are placed my contributions in the development of higher order RD schemes, in collaboration with my colleagues at INRIA and at the von Karman Institute.

As recalled in section §2.5, some initial work in this direction is discussed in [11] and in [54], based on two different strategies. Both strategies are inspired by the error analysis of section §2.2.2, and in particular by the consistency estimate (2.24) and by propositions 2.2.5 and 2.2.7. The methods proposed in [11, 54] consider higher polynomial approximations based either on the use of higher degree  $P^k$  Lagrange finite element [11], or on the reconstruction of nodal gradients to be used in the evaluation of the face flux integrals [54].

The developments described in the following paragraphs follow [11]. In particular, the principle followed is to increase  $k$  in (3.11), trying to minimize the increase in  $n_{\text{dof}}$ . The way in which this is done is by sticking to a continuous interpolation as in [11]. The aim of this work is to embed the schemes with some form of nonlinear discontinuity capturing, in the same spirit of what is done for second order schemes, as discussed in section §2.2.5. We will first recall the construction of higher order multidimensional upwind with crosswind dissipation presented in [RVAD05, VRD06, ARN<sup>+</sup>06] (see also [RVAD08, VQRD11], and the PhD thesis [261]), and then describe the non-upwind approach, based on higher order nonlinear variants of the LF scheme, proposed in [ALRT09, ABR11, ALR11] (see also [ALR08a, ALR10] and [ARN<sup>+</sup>06, ARTL07, ALR08b, ALR08a, LAR09, RAA09]).

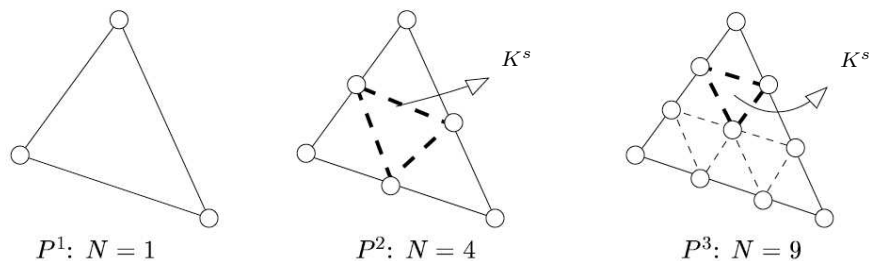


Figure 3.6: Sub triangulations of  $P^k$  triangles

<sup>2</sup>[http://ec.europa.eu/research/transport/projects/items/adigma\\_en.htm](http://ec.europa.eu/research/transport/projects/items/adigma_en.htm)

### 3.2.2 Multidimensional upwind schemes

In this section, we focus on the scalar case. A first approach to construct higher order schemes with some shock capturing capabilities has consisted in making use of relation (3.7). The starting point is given by the schemes proposed in [11]. These schemes make use of higher order  $P^k$  Lagrange elements. Each element is sub-divided into sub-elements by using the conformal  $P^1$  sub-triangulation defined by all the degrees of freedom, as shown on figure 3.6. An approximation of the steady solution of (2.1) is sought by (cf. figure 3.6 for the notation)

1.  $\forall K \in \Omega_h : \forall K^s \in K$  compute  $\phi^{K^s} = \int_{K^s} \nabla \cdot \mathcal{F}_h$
2.  $\forall K \in \Omega_h$  : distribute the sub-elemental residuals. Let  $\phi_j^{K^s}$  be the amount of fluctuation sent to  $j \in K^s$  :  $\sum_{j \in K^s} \phi_j^{K^s} = \phi^{K^s}$
3.  $\forall i \in \Omega_h$  : assemble contributions from all  $K^s \in K_j$  and evolve according to

$$|C_i| \frac{u_i^n - u_i}{\Delta t} = - \sum_{K \in K_i} \sum_{\substack{K^s \in K \\ i \in K^s}} \phi_i^{K^s} \quad (3.13)$$

where the dual cell  $C_i$  is defined on the  $P^1$  sub-triangulation as shown on figure 3.7.

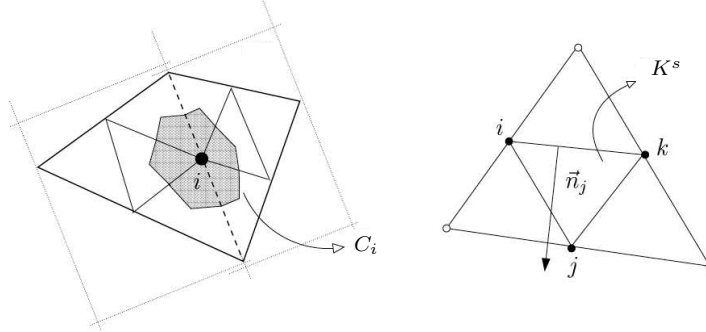


Figure 3.7: Median dual cells and normals for sub-triangulated  $P^k$  triangles

The criteria for the distribution are similar to those discussed in paragraph 3 :

**Accuracy preservation** An *Accuracy preserving* scheme is obtained for  $\phi_i^{K^s} = \beta_i^{K^s} \phi^{K^s}$ .

As a consequence of proposition 2.2.7 and of the consistency estimate of lemma 2.2.8, accuracy preserving schemes have a  $\mathcal{O}(h^{k+1})$  truncation error on  $P^k$  elements ;

**Positivity** A positive scheme is one for which  $\phi_i^{K^s} = \sum_{j \in K} c_{ij} (u_i - u_j)$  with  $c_{ij} \geq 0$ . Positive schemes enjoy a discrete maximum principle (proposition 2.2.3).

**Multidimensional upwinding** The  $P^1$  definition of upstream and downstream geometrical entities is retained locally. In particular, on every sub-element one defines nodal normals  $\vec{n}_j$  (as on the right picture on figure 3.7), and

$$k_j = \vec{a}_{K^s} \cdot \vec{n}_j \quad (3.14)$$

with  $\vec{a}_{K^s}$  a sub-elemental average of the flux Jacobian (2.2). Multidimensional Upwind schemes are those for which :  $k_j \leq 0 \Rightarrow \phi_j^{K^s} = 0$ .

As shortly discussed in section §2.5, the scheme proposed in [11] fits into this framework. In particular, in every  $K^s$ , the authors suggest the following nonlinear construction :

1. Compute a  $P^1$  N scheme using the three nodes of  $K^s$ . Let  $\phi_i^N$  denote on  $K^s$  the split residual for node  $i$  ;
2. Apply a limiter to  $\{\phi_i^N\}_{i \in K^s}$  obtaining  $\forall i \in K^s$

$$\beta_i^{K^s} = \frac{\max(0, \phi_i^N \phi^{K^s})}{\sum_{j \in K^s} \max(0, \phi_j^N \phi^{K^s})}$$

3. Set  $\phi_i^{K^s} = \beta_i^{K^s} \phi^{K^s}$

Note that the  $\phi^{K^s}$  uses the higher order interpolation, hence all the information on  $K$ , while the N scheme residuals only use the three nodes of  $K^s$ . In particular,

$$\sum_{j \in K^s} \phi_j^N = \phi^{P^1} \neq \phi^{K^s} \quad (3.15)$$

In [RCD05, Ric05], the well-posedness of nonlinear schemes based on the use of limiters is analyzed. As a particular case of a more general result, we can prove the following.

**Proposition 3.2.1** (Well posedness of the limiting). *Given a high order fluctuation  $\phi^h$ , and a set of positive low order split residuals  $\phi_j^P$  such that  $\sum_j \phi_j^P = \phi^1$ . A sufficient condition for the limiter*

$$\beta_i = \frac{\max(0, \phi_i^P \phi^h)}{\sum_{j \in K^s} \max(0, \phi_j^P \phi^h)} \quad (3.16)$$

to yield a consistent scheme is that  $\phi^1 \phi^h > 0$ .

*Proof.* Let  $\phi^h \neq 0$ , and set  $x_j = \phi_j^P \phi^h$ . In order for (3.16) to yield at least one non-null coefficient, there must be at least one positive  $x_j$ . Consider now

$$\sum_j x_j = \phi^1 \phi^h$$

If  $\phi^1 \phi^h$  is strictly positive, then there must be at least one positive  $x_j$ . Thus, by construction, the limiter will yield a consistent scheme, that is a scheme for which

$$\sum_j \beta_j \phi^h = \phi^h \quad \text{and} \quad \sum_j \beta_j = 1$$

If instead  $\phi^1 \phi^h \leq 0$ , there is no assurance that  $x_j > 0$  for at least one  $j$ . In particular, if  $x_j \leq 0 \forall j$ , the limiter will give an inconsistent discretization, for which  $\beta_j = 0 \forall j$   $\square$

In the case of the scheme proposed by [11], there is no *a priori* assurance that  $\phi^{P^1} \phi^{K^s} > 0$  on every sub-element  $K^s$ . In particular, the lack of control on the values of the degrees of freedom outside  $K^s$  make it quite easy to find counterexamples where at least  $\phi^{P^1} \phi^{K^s} = 0$ .

This explains why in [11] the authors mention that “It seems more important for higher order schemes than for the second order PSI scheme that  $\sum_j \hat{\beta}_j = 1$  exactly. We have chosen to compute the revised weights (in pseudo fortran) as”

$$\beta_i = \frac{\max(0, \phi_i^P \phi^h) + 1.e^{-10}}{\sum_{j \in K^s} \max(0, \phi_j^P \phi^h) + 3.e^{-10}}$$

The small correction added in the limiter cures the inconsistency foreseen by proposition 3.2.1 by reverting to a central distribution.

To overcome this problem, we proposed a simple idea in [RVAD05, VRD06] : define sub-elemental linear schemes starting from the LDA and adding a certain amount of the crosswind dissipation term (3.7). We define on  $K^s \in K$  (cf. figure 3.7 and (3.14)) :

**LDA** A sub-element LDA scheme, defined as in [11]

$$\phi_i^{\text{LDA}} = k_i^+ N^{-1} \phi^{K^s}, \quad N = \sum_{j \in K^s} k_j^+$$

**N scheme** A sub-element linear scheme, defined based on relation (3.7) :

$$\phi_i^{\text{N}} = \phi_i^{\text{LDA}} + d_i^{K^s}, \quad d_i^{K^s} = \sum_{j \in K^s} k_i^+ N^{-1} k_j^+ (u_i - u_j)$$

**B scheme** A blended scheme (cf. section §2.2.6)

$$\phi_i^{\text{B}} = l(u_h) \phi_i^{\text{N}} + (1 - l(u_h)) \phi_i^{\text{LDA}} = \phi_i^{\text{LDA}} + l(u_h) d_i^{K^s}$$

where, using (3.7), we have recast the B scheme as the LDA plus a nonlinear crosswind dissipation term ;

**LN scheme** Obtained by applying (3.16) to the sub-elemental N scheme defined above.

In this case  $\sum_j \phi_j^{\text{N}} = \phi^{K^s}$ , so the limiting is always well-posed. However, the definition of the  $d_i^{K^s}$  term does not account for all the information contained in  $\phi^{K^s}$ . In particular, nodes that do not belong to  $K^s$  are not included in this term, while contributing to the value of  $\phi^{K^s}$ . As a consequence, the N scheme itself is in general not positive.

To check how much this affects the results, we consider two tests. The first involves a linear flux  $\mathcal{F} = \vec{a}u$  with the “solid body rotation” advection speed  $\vec{a} = (y, -x)$ . In the spatial domain  $[-1, 1] \times [0, 1]$  we solve the steady problem with boundary condition

$$\begin{cases} u(x < 0, 0) = 1 & \text{if } -0.7 \leq x \leq -0.3 \\ u(x < 0, 0) = 0 & \text{otherwise} \\ u(1, x > 0) = u(0, y) = 0 \end{cases}$$

Steady solutions are computed on an unstructured triangulation with the topology on the left on figure 3.8. The  $P^1$  computations are run on the conformal sub-triangulation so that  $P^1$  and  $P^2$  results are obtained with the exact same number of degrees of freedom. The size of the conformally refined  $P^1$  mesh is  $h = 1/40$ . The iterative convergence of the schemes is very fast, as shown on the right on figure 3.8.



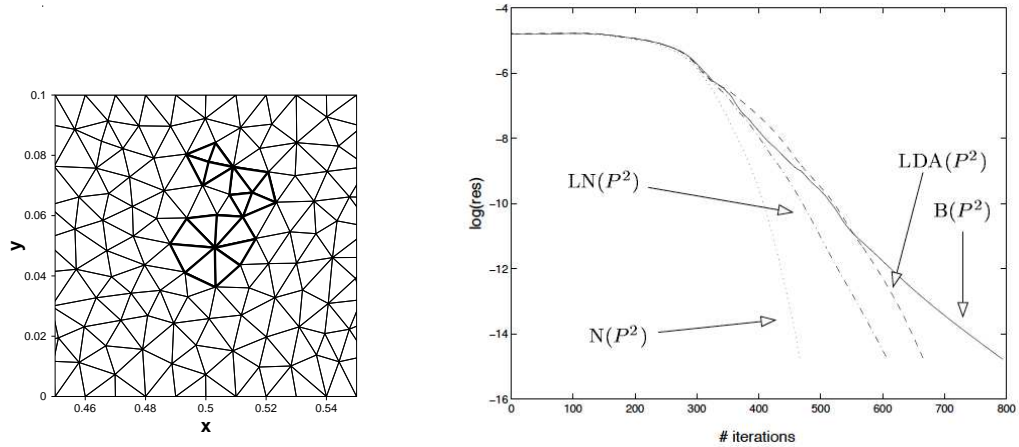


Figure 3.8: Unstructured mesh and iterative convergence for the rotation of a top-hat.

Contours of the steady solutions are reported on figure 3.9. As expected, the shear is thinner with the nonlinear  $P^2$  schemes (lower pictures, left column) than with the blended  $P^1$  scheme (left column second from top). Oscillations are within 3% for the  $B(P^2)$  scheme, and within 9% for the  $LN(P^2)$  scheme, so even though they seem indistinguishable, the two solutions do present important differences. The outlet profiles on the right in the same figure confirm the monotonicity of the discrete solution.

A very similar behavior is observed when considering the nonlinear problem obtained with the choice of the flux  $\mathcal{F} = (e^u, u)$ . On the spatial domain  $[-0.025, 1.2] \times [0, 0.5]$  we solve the associated conservation law with boundary conditions

$$\begin{cases} u(x, 0) = \sin(2\pi x) & \text{if } 0 \leq x \leq 1 \\ u(x, 0) = 0 & \text{otherwise} \\ u(-0.25, y) = 0 \end{cases}$$

Steady solutions have been approximated on an unstructured triangulation with the topology shown on the left on figure 3.8. Computations with  $P^1$  schemes on the conformally refined triangulation have been performed for comparison. The mesh size of the refined triangulation is  $h = 1/40$ . The results are reported on figure 3.10. The solution contours on the left column show a very oscillatory behavior for the LDA scheme (top picture), and monotone shock capturing for the B schemes (middle and bottom). Among these, The  $B(P^1)$  result (middle) shows a slightly better capturing of the shock. However, the oscillations obtained with the  $B(P^2)$  schemes are below 10%. The right column shows that again the iterative convergence of all the schemes is quite fast, and that indeed the oscillations of the  $LDA(P^2)$  scheme are considerably reduced by the blending. The  $LN(P^2)$  results are again very close, the oscillations being gain slightly more pronounced.

The accuracy of the schemes is easily confirmed by grid convergence studies [RVAD05, VRD06], showing the  $k + 1$ -th order of accuracy of the LDA and B and LN schemes in the  $P^1$ ,  $P^2$ , and  $P^3$  case. A more interesting test is the following : we consider again the “solid body rotation” advection speed  $\vec{a} = (y, -x)$ . In the spatial domain  $[-1, 1] \times [0, 1]$  we solve

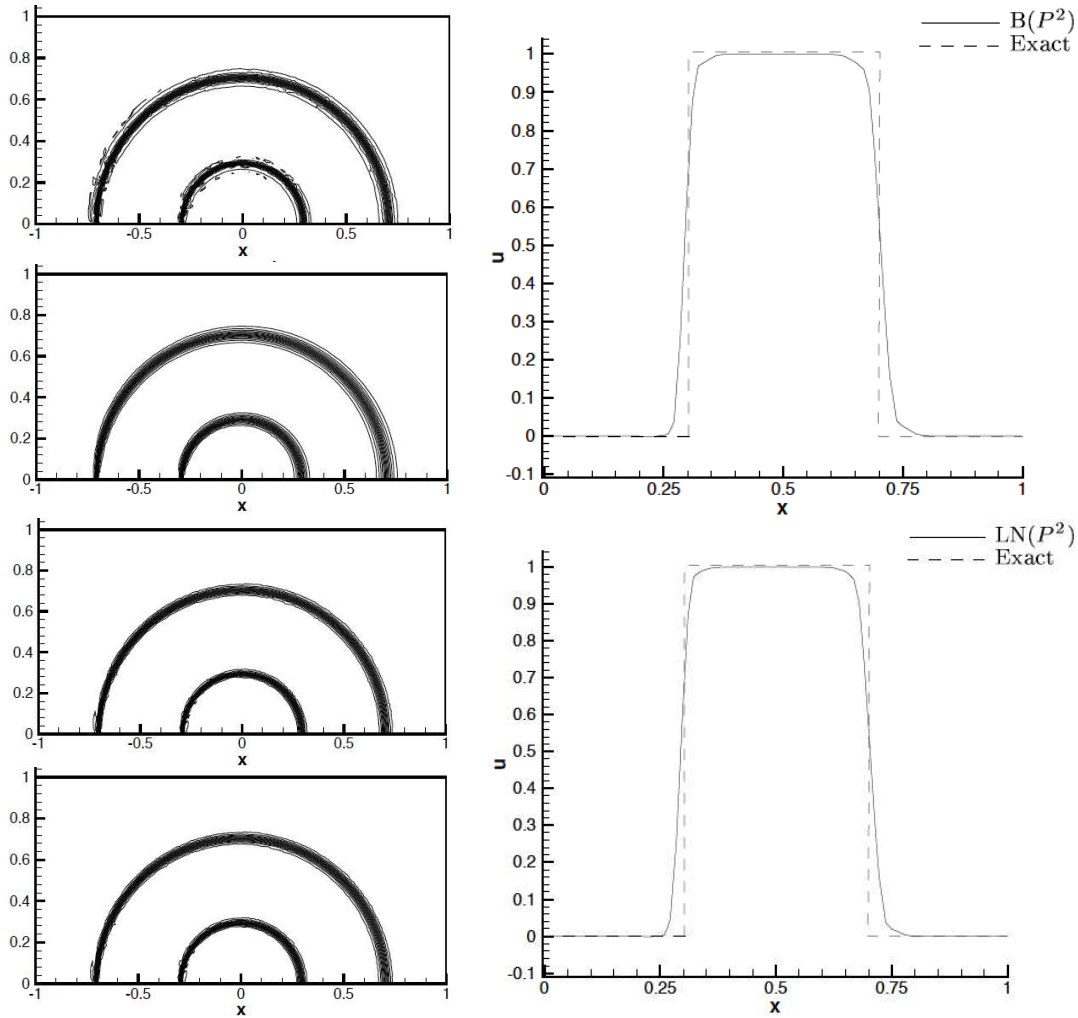


Figure 3.9: Rotation of a top-hat. Left : solution contours (From the top : LDA( $P^2$ ) - B( $P^1$ ) - B( $P^2$ ) - LN( $P^2$ )). Right : outlet profiles (Top : B( $P^2$ ) - bottom : LN( $P^2$ ))

the steady problem with boundary condition

$$\begin{cases} u(x < 0, 0) = \sin(10\pi x) \\ u(1, x > 0) = 0 \\ u(0, y) = 0 \end{cases}$$

Third order results are compared with the second order ones on the  $P^1$  conformally refined triangulations (comparison for the same number of degrees of freedom). The size of the refined mesh is ( $h = 1/40$ , 8 cells per period). The iterative convergence is again very rapid (not shown). We only report the results of the B schemes, on figure 3.11. These results show the incredible error reduction brought by the increase in polynomial approximation.

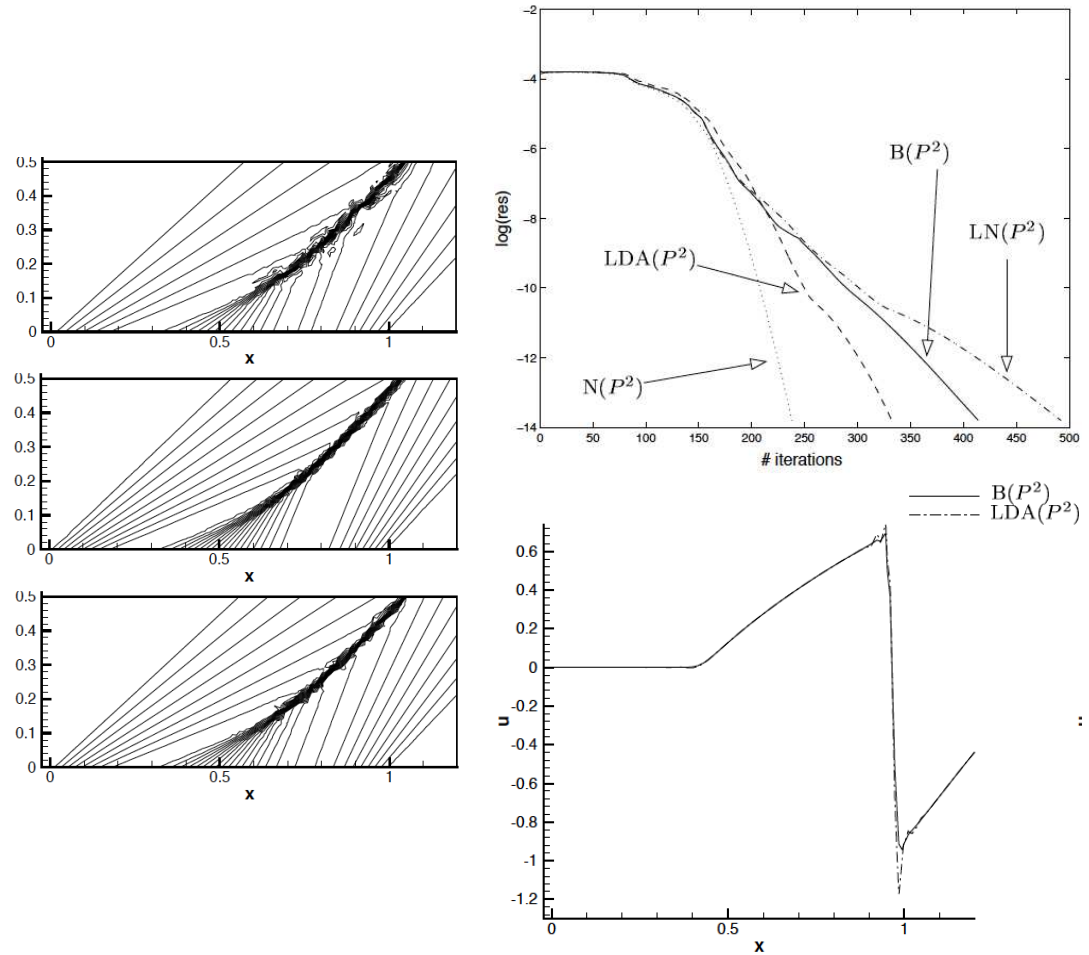


Figure 3.10: Conservation law with an exponential flux. Left : solution contours (From the top :  $LDA(P^2) - B(P^1) - B(P^2)$ ). Right : iterative convergence and outlet profiles

The benefit of this construction is related to the fast iterative convergence (at least for scalar problems), and to the fact of having constructed a simple non-oscillatory third order scheme of the type “high order plus cross-wind dissipation”. Unfortunately, things start getting a little worse in the  $P^3$  case, when the control of the oscillations associated to the local cross-wind dissipation is weaker. The extension to the Euler equations is discussed in the PhD of Villedieu [261], and it confirms these observations for transonic and mildly supersonic problems. Further extensions to advection diffusion and to the laminar Navier-Stokes equations are discussed in [RVAD05, VRD06, RVAD08, VQRD11].

The limitations mentioned justify the investigation of different approaches.

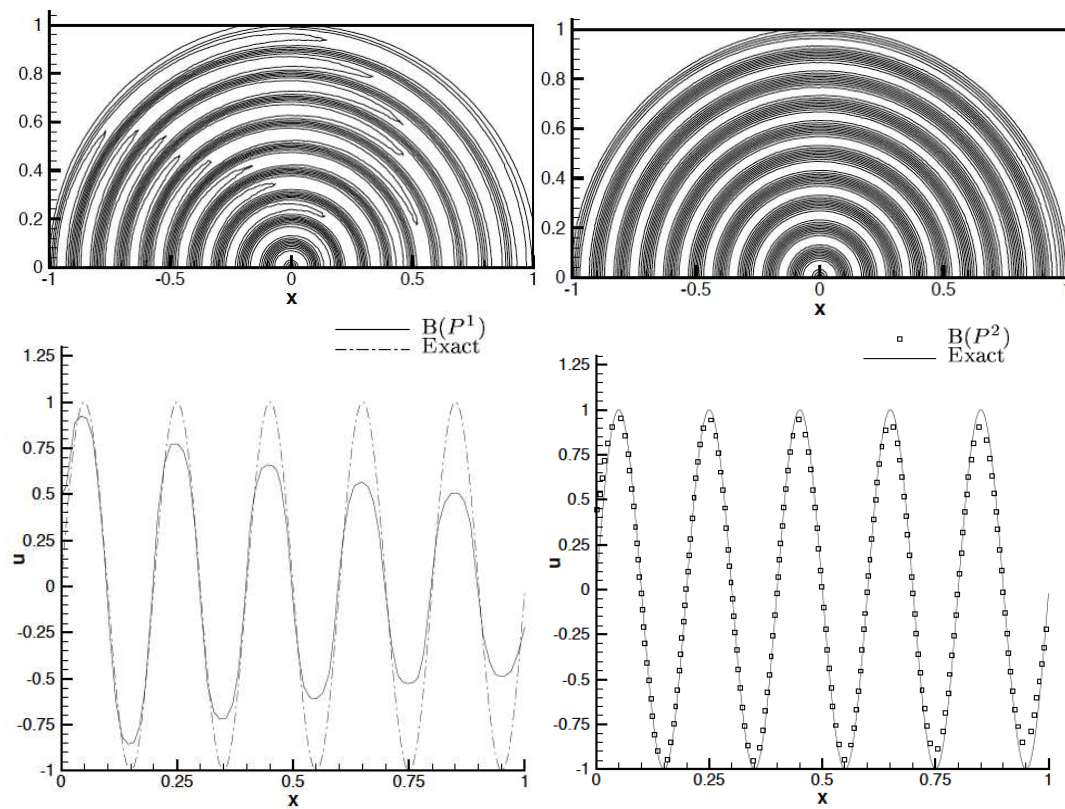
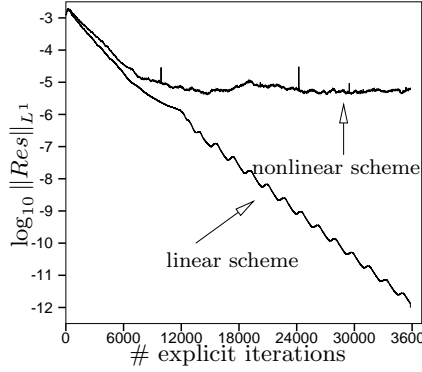


Figure 3.11: Rotation of a high frequency sinusoidal :  $B(P^1)$  (left) and  $B(P^2)$  results.

### 3.2.3 Non-upwind higher order schemes

#### Upwinding, stability and systems

The work that has led to the construction object of this paragraph are motivated by the poor iterative convergence of the matrix system extension of nonlinear RD schemes<sup>3</sup>.



Differently from what we have observed in the previous section, the typical iterative behavior of these matrix RD is the one depicted on the left [1, CRD02, 96, 2]. It is this flaw that has given motivation to investigate simpler non-upwind discretizations, *e.g.* based on the LF scheme (2.39) [10, 180]. Even in the scalar case, when trying to make a PSI scheme starting from (2.39), one obtains a nonlinear scheme with an iterative convergence surprisingly close to the one shown on the left [10, 180]. In both cases, the consequence of the lack of iterative convergence is that mesh convergence is polluted by the remainder in the iterative procedure (2.11), and second order of accuracy is often not observed in practice [96, 95].

This flaw is related to a stability problem which is not observed for MU schemes in the scalar case. In this case, a simple energy balance can be obtained as (cf. equation (2.11))

$$\sum_{i \in \Omega_h} u_i |C_i| \frac{du_i}{dt} = \frac{d}{dt} \int_{\Omega_h} \mathcal{I}_h = - \sum_{K \in \Omega_h} \overbrace{\sum_{j \in K} u_j \phi_j^K}^{\phi_K^\mathcal{E}}$$

having denoted by  $\mathcal{I}$  the energy density  $u^2/2$ , by  $\mathcal{I}_h$  its piecewise polynomial approximation (note that  $\mathcal{I}_h \neq \mathcal{I}(u_h)$ ), and by  $\phi_K^\mathcal{E}$  the local energy production of the discretization. For scalar advection, in [RVAD05, Ric05, DR07, 6] it is shown that in the  $P^1$  case

**Proposition 3.2.2** (MU schemes and NRG balance).

1. in one target elements all MU schemes are locally dissipative

$$\phi_{MU}^\mathcal{E} = \oint_{\partial K} \mathcal{I}_h \vec{a} \cdot \vec{n} + \epsilon_{MU}, \quad \epsilon_{MU} \geq 0$$

2. in 2 target elements the energy budget of the N and LDA schemes is (cf. equation (3.9))

$$\phi_N^\mathcal{E} = \oint_{\partial K} \mathcal{I}_h \vec{a} \cdot \vec{n} + \sum_{i,j \in K} u_i M_{ij}^N u_j$$

with  $M_{ij}^N$  positive semi-definite, and (cf. equation (2.30))

$$\phi_{LDA}^\mathcal{E} = N \left( \mathcal{I}(u_{out}) - \mathcal{I}(u_{in}) \right) + \frac{1}{2} (u_{out} - u_{in}) N (u_{out} - u_{in})$$

<sup>3</sup>the only exception is represented by computations of supersonic flows, when using the PSI scheme on the decoupled scalar supersonic invariants [201, 229, 124]

This means that the linear N and LDA schemes verify the energy balance [RVAD05, Ric05, DR07]

$$\frac{d}{dt} \int_{\Omega_h} \mathcal{I}_h = - \sum_{K \in \Omega_h} \Delta_{\bar{a}} \mathcal{I}_h - \sum_{K \in \Omega_h} \delta \mathcal{E}, \quad \delta \mathcal{E} \geq 0$$

the first term on the right hand side representing a local net energy flux, the second a dissipation term. The analysis does generalize to (symmetrizable) systems [DR07, Ric05, 6].

For the PSI scheme (or limited N scheme), we are left with point 1. of the proposition which does show the existence of a dissipative mechanism related to upwinding, as confirmed by all practical observations showing a rapid and monotone iterative convergence, and second order rates when refining the mesh (for smooth solutions). When passing to systems, this geometrical mechanism disappears, and, unless some scalar decomposition is introduced (cf. section §2.3.2) the concept of 1 or 2 target elements does not make any sense anymore. The benefit of multidimensional upwinding is definitely lost when some form of matrix limiter is applied to the N scheme [96, DR07, 2].

The same behavior is observed in the scalar case when applying limiter (2.43) (or equivalently (2.44)) to the LF scheme to produce a nonlinear high order splitting. If however in one target cells, and only in these, this limited LF scheme is replaced by the MU scheme distributing everything to the only downstream node, thus restoring the effects of point 1. in proposition 3.2.2, the problem is cured and not only the resulting scheme is convergent but also high order [180].

An algebraic view of this phenomenon is presented in [2], where a practical solution is proposed. This solution constitutes the basis for many of the developments of this manuscript, and in particular for those discussed in this paragraph. The solution proposed is based on the following experimental observations :

- Even though not converging to machine accuracy, the limited schemes are consistent, as shown by the fact that grid convergence is observed in practice, even though often only with first order rates ;
- The lack of convergence is related to mild spurious modes not detected by the scheme, however bounded due to the effects of the underlying positivity preserving procedure ;
- These modes are especially relevant in smooth regions.

The solution proposed in [2] is to add in smooth regions a dissipative term that retains the residual character of the approach. In particular, the author proposes to add the Streamline Dissipation term (cf. equation (2.35)), so that the final scheme reads

**Definition 3.2.3** (Limited and Stabilized Positive RD -  $P^1$  case ).  $\forall K \in \Omega_h$ , given the element residual  $\phi^K$  and a set of positivity preserving first order linear distributed residuals  $\phi_i^P$ , such that  $\sum_j \phi_j^P = \phi^K$ , a high order, limited and stabilized RD scheme is defined as :

1. Apply the linearity preserving limiter :  $\phi_i^* = \beta_i^* \phi^K = \frac{\max(0, \phi_i^P \phi^K)}{\sum_{j \in K} \max(0, \phi_j^P \phi^K)} \phi^K$
2. Evaluate a smoothness sensor  $\delta(u_h)$  :  $\delta = C_u h$  in discontinuities, and  $\delta = 1$  elsewhere ;

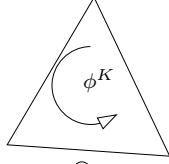
3. Add the Streamline dissipation :  $\phi_i^K = \phi_i^* + \delta(u_h)\beta_i^{SD}\phi^K = \left(\beta_i^* + \delta(u_h)\frac{k_i}{|K|}\tau\right)\phi^K$

The generalization of this construction to arbitrary elements is described in [ALRT09, ABJR11, ALR11] (see also [ALR08a, ALR10] and [ARN<sup>+</sup>06, ARTL07, ALR08b, ALR08a, LAR09, RAA09]), and recalled in the following subsection.

### A Limited LF schemes for general elements

The starting point of the construction is a high order continuous Lagrange (or other [13]) polynomial approximation. Let us consider standard  $P^k$  and  $Q^k$  Lagrange elements [286]. Using the result of proposition 2.2.7 and the estimate of lemma 2.2.8, we can formulate the following systematic construction of a *formally*  $k + 1$ th order positive scheme for (2.1) :

1.  $\forall K \in \Omega_h$  compute  $\phi^K = \int_K \nabla \cdot \mathcal{F}_h(u_h)$  ;

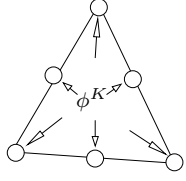


2.  $\forall K \in \Omega_h$  compute  $\forall i \in K$  the first order Lax-Friedrich's distribution

$$\phi_i^{LF} = \frac{1}{C_K} \left( \phi^K + \alpha \sum_{j \in K} (u_i - u_j) \right), \quad \alpha > |K| \sup_K |\vec{a}(u_h) \cdot \nabla \varphi_j|, \forall j \in K$$

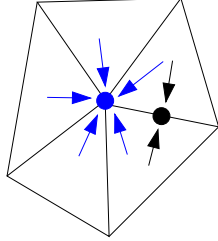
$C_K$  denoting the number of degrees of freedom in  $K$  ;

3.  $\forall K \in \Omega_h$  compute  $\forall i \in K$  the limited LF splitting



$$\phi_i^K = \phi_i^* = \frac{\max(0, \phi_i^{LF} \phi^K)}{\sum_{j \in K} \max(0, \phi_j^{LF} \phi^K)} \phi^K$$

4.  $\forall i \in \Omega_h$  assemble the signals from surrounding elements and evolve nodal values



$$u_i^{n+1} = u_i^n - \omega_i \sum_{K \in K_i} \phi_i^K$$

with  $\omega_i$  an iteration parameter.

The procedure is now applied to two scalar problems. First, we seek a steady solution to the Burger's equation, obtained setting  $\mathcal{F} = (u^2/2, u)$  in (2.1), on the spatial domain  $[0, 1]^2$  with boundary conditions

$$\begin{cases} u(x, 0) = 1.5 - 2x \\ u(0, y) = 1.5 \end{cases}$$

Then, we solve the solid body rotation problem, obtained setting  $\mathcal{F} = (yu, -xu)$  in (2.1), on the spatial domain  $[-1, 1] \times [0, 1]$  with boundary conditions

$$\begin{cases} u(x < 0, 0) = \sin^2(2\pi x) & \text{if } -0.75 \leq x \leq -0.25 \\ u(x < 0, 0) = 0 & \text{otherwise} \\ u(-1, y) = 0 \\ u(x > 0, 1) = 0 \end{cases}$$

Both problems are solved on an unstructured triangulation with the topology shown on the left on figure 3.8 and  $h = 1/40$  with the  $P^1$  scheme and with the  $P^2$  scheme. The results are summarized in figures 3.12 and 3.13.

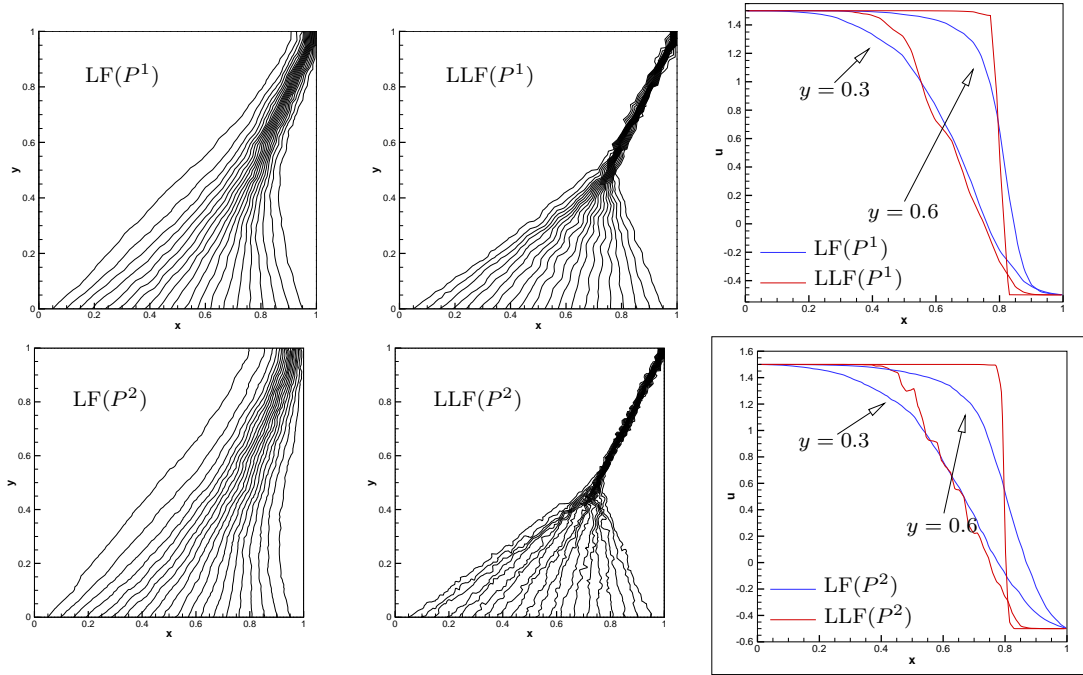


Figure 3.12: Burger's equation :  $P^1$  and  $P^2$  LF and LLF results

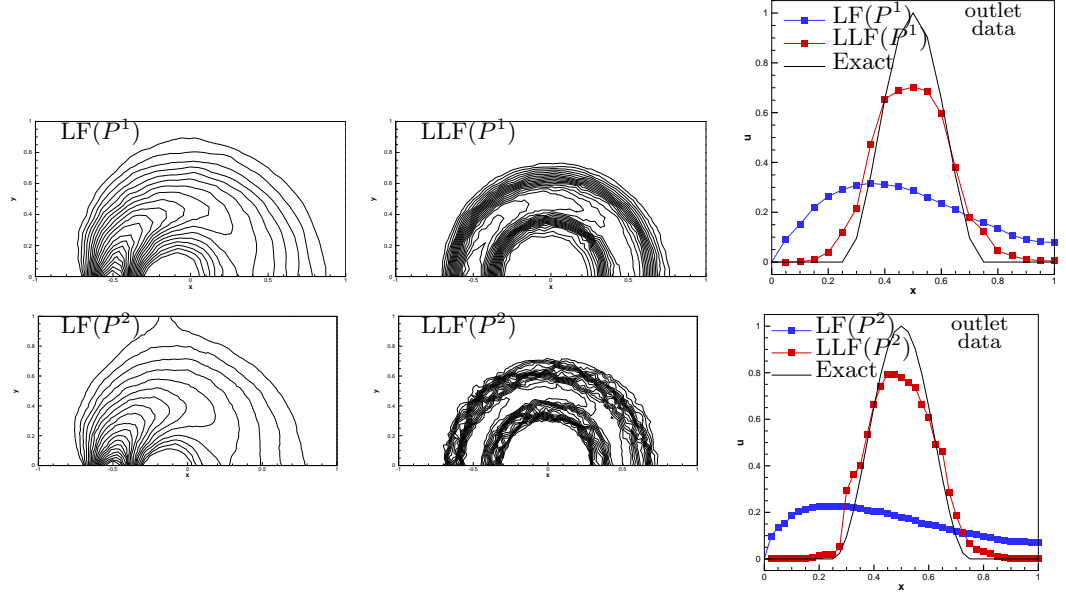
Several remarks can be made. Let us start with the Burger's results on figure 3.12. The difference between the first order LF and the high order variants is impressive. The shock in the Burger's problem is captured perfectly, without any oscillation. The width of the shock is one element in both the  $P^1$  and  $P^2$  case, hence the  $P^2$  shock layer contains more degrees of freedom. However, the linear part of the solution before the shock foot is noisy, especially in the  $P^2$  solution. This suggests that the spatial frequency of this spurious mode is the distance between two degrees of freedom.

The results for the rotation problem are quite similar. The resolution of the limited schemes is way better than that of the linear one. However, unphysical plateaus appear in the high order solutions. This is especially visible in the outlet profile of the  $P^2$  scheme, even though a less pronounced plateau is also visible in the  $P^1$  outlet data ( $x \approx 0.75$ .)

A grid convergence study shows that all the nonlinear results are only first order accurate. Something is missing. As anticipated, the problem is related to a lack of numerical dissipation. To cure this, we could naively add an extra term as in definition 3.2.3, and correct the scheme using some formal generalization of the SD term

$$\phi_i^K = \phi_i^* + \frac{1}{|K|} \delta(u_h) \left( \int_K \tilde{a}(u_h) \cdot \nabla \varphi_i \right) \tau \phi^K$$



Figure 3.13: Rotation of a  $\cos^2$  profile :  $P^1$  and  $P^2$  LF and LLF results

One might even argue that the local energy budget of the scheme is now

$$\sum_{i \in K} u_i \phi_i^K = \sum_{i \in K} u_i \phi_i^* + \frac{1}{|K|} \delta(u_h) \left( \int_K \vec{a}(u_h) \cdot \nabla u_h \right) \tau \phi^K = \sum_{i \in K} u_i \phi_i^* + \overbrace{\frac{\tau \delta(u_h)}{|K|} (\phi^K)^2}^{\geq 0}$$

so that we have apparently added dissipation to the scheme. If one does that, the problem solved only in the  $P^1$  case, but the noise and the plateaux in the  $P^2$  results remain.

This leads to another pitfall of the residual approach, as initially formulated.

**Proposition 3.2.4** (Fall of the  $\beta\phi$  paradigm, 2d advection). *Consider the solution of*

$$\vec{a} \cdot \nabla u = 0$$

with  $\vec{a}$  constant. Any scheme of the form

$$0 = \sum_{K \in \Omega_h} \beta_i^K \phi^K = \sum_{K \in \Omega_h} \beta_i^K \int_{\partial K} u_h \vec{a} \cdot \vec{n}, \quad \forall i \in \Omega_h$$

cannot be freed of high frequency spurious modes whatever the form of  $\beta_i^K$ , if  $K$  is a  $P^k$  triangle with  $k > 2$  and if  $K$  is a  $Q^k$  quadrilateral  $\forall k \geq 1$ .

*Proof.* We start showing the existence of at least one spurious mode (independent of  $\vec{a}$ ) for all elements  $P^k$  and  $Q^k$ ,  $k \geq 2$ . Denote by  $C_f$  the number of degrees of freedom on each face  $f \in \partial K$  minus 2. We consider the mode defined by  $u_j = 1$  if  $j$  is a vertex, otherwise on each

$f \in \partial K$  we set  $\forall j \neq v$

$$u_j = -\frac{2}{C_f} \frac{\int_f \varphi_v}{\int_f \varphi_j}$$

having denoted with  $v$  one of the two vertices forming face  $f$ . The mode is compatible with the continuity of the representation, and with the adoption of hybrid meshes. For  $P^k$  elements with  $k \geq 3$  and  $Q^k$  elements with  $k \geq 2$ , the value of the solution at nodes in the elements remains arbitrary. For this mode, one easily checks that  $\phi^K = 0, \forall K$ , so that any scheme of the type  $\beta\phi$  will preserve it forever.

The only remaining element is the  $Q^1$  quadrilateral which is easily checked to suffer from the checkerboard spurious mode in which  $u$  oscillates between  $-1$  and  $1$  on every face. In this case as well  $\phi^K$  vanishes identically.  $\square$

A similar property can be easily derived in the three dimensional case as well. The main problem is how to define the spurious mode in meshes with elements with different face topologies (*e.g.*  $P^k \times Q^k$  prismatic elements) while ensuring that the mode is compatible with the continuity of the approximation. Details are left out of this manuscript.

The important consequence of proposition 3.2.4 is that we have to start looking for schemes exploiting the sub-elemental variation of the discrete solution. For this reason, in [ALRT09, ALR11] the approach of [2] is generalized as follows

**Definition 3.2.5** (Limited and Stabilized Positive RD - general case).  $\forall K \in \Omega_h$ , given  $\phi^K$  and a set of positivity preserving first order linear distributed residuals  $\phi_i^P$  :

1. Apply the linearity preserving limiter :

$$\phi_i^* = \beta_i^* \phi^K = \frac{\max(0, \phi_i^P \phi^K)}{\sum_{j \in K} \max(0, \phi_j^P \phi^K)} \phi^K$$

2. Evaluate a smoothness sensor  $\delta(u_h) : \delta = C_u h$  in discontinuities, and  $\delta = 1$  elsewhere ;
3. Add the Streamline dissipation :

$$\phi_i^K = \phi_i^* + \delta(u_h) \phi_i^{SD} = \beta_i^* \phi^K + \delta(u_h) \int_K (\vec{a}(u) \cdot \nabla \varphi_i) \tau(\vec{a}(u) \cdot \nabla u_h)$$

**Remark 3.2.6** (Conservation and quadrature). *Not surprisingly, the additional term IS the streamline dissipation term of Hughes and co-workers [146]. However, we will give several remarks that allow considerable simplification.*

**Discrete conservation** *Discrete conservation is guaranteed already by the nonlinear term, and the SD term does not affect that since, due to (2.6)*

$$\sum_{j \in K} \phi_j^{SD} = \int_K (\vec{a}(u) \cdot \sum_{j \in K} \nabla \varphi_j) \tau(\vec{a}(u) \cdot \nabla u_h) = 0$$

*this means that we can freely use the quasi-linear (or any other simplified) form of the equations without having to worry about discrete conservation ;*

**Discrete form** In practice, the extra term is replaced by its evaluation is a certain number of quadrature points so that the relevant term to analyze is

$$\phi_i^{SD_h} = \sum_q \omega_q |K| \bar{a}(u_h(x_q)) \cdot \nabla \varphi_i(x_q) \tau \bar{a}(u_h(x_q)) \cdot \nabla u_h(x_q)$$

having assumed a constant value for  $\tau$  over the element. Note, that the remark on conservation apply to  $\phi_i^{SD_h}$  as well ;

**Accuracy** It is easily shown that, if  $u$  is a sufficiently smooth exact solution, then [63, 105]  $\nabla(u_h - u) = \mathcal{O}(h^k)$ . This immediately allows to prove that with all standard definitions of  $\tau$  one has  $\phi_i^{SD_h}(u_h) = \mathcal{O}(h^{k+2})$  whenever  $u_h$  is the approximation of a regular enough classical solution. This means that the addition of this term does not pollute the formal accuracy since we verify the conditions of proposition 2.2.7.

**Practical evaluation and dissipative character** How many quadrature points should one use ? Exact integration being impractical for complex definitions of  $\mathcal{F}$ , we can answer this question generalizing the reasoning of [ALRT09]. Everything boils down to the relation  $\mathcal{F}(u)$ , and to the energy/entropy budget associated to the extra term. In the scalar case, assuming that  $u^2$  is an entropy for our problem, this budget is

$$\sum_{j \in K} u_j \phi_i^{SD_h} = \sum_q \omega_q |K| \tau (\bar{a}(u_h(x_q)) \cdot \nabla u_h(x_q))^2$$

which is positive definite, unless  $\bar{a}(u_h(x_q)) \cdot \nabla u_h(x_q) = 0 \forall q$ . In order to correctly reproduce the behavior of the discrete solution, it is suggested in [ALRT09] that the number and location of evaluation points should be the minimum guaranteeing that the polynomial  $\bar{a}(u_h(x_q)) \cdot \nabla u_h(x_q)$  is uniquely defined. For example, for constant  $\bar{a}$ , it is suggested to use one point in the  $P^1$  case, 3 non co-linear points (e.g the vertices) in the  $P^2$  case, 4 points in the  $Q^1$  case (the minimum is 2 but 4 points allows a symmetric formula), etc. etc.

In general, this choice will depend on the relation  $\mathcal{F}(u)$ , or better  $\mathcal{F}(v)$ ,  $v$  being the entropy variable  $v = \partial_u s(u)$ , with  $s(u)$  the entropy. For example, if this relation can be modeled with a polynomial of degree  $p_s$ , then for a  $P^k$  approximation the choice of the evaluation points should be such that a polynomial of degree  $(p_s + 1)k - 1$  over the element is uniquely defined.

**Choice of the weights** Whatever the choice of the weights, provided that the evaluation points are correctly chosen, then the scheme is conservative, it verify the conditions for  $k + 1$ th order of accuracy, and the discrete SD term will be dissipative. The weights can be set to an arbitrary value. In all the results that follow  $\omega_q = 1/C_K, \forall q$ , with  $C_K$  the number of degrees of freedom of the element.

**Smoothness sensor** The best definition for  $\delta(u_h)$  found so far is

$$\delta(u_h) = 1 - \max_{j \in K} \max_{K' \in K_j} \max_{l \in K'} \frac{|u_l - \bar{u}_{K'}|}{|u_l| + |\bar{u}_{K'}| + 1.e^{-10}}$$

having denoted with  $\bar{u}$  the arithmetic average of the nodal values. Details on a smart, compact, implementation of this formula that profits of the iterative strategy used to solve the discrete equations are given in [ALR11].

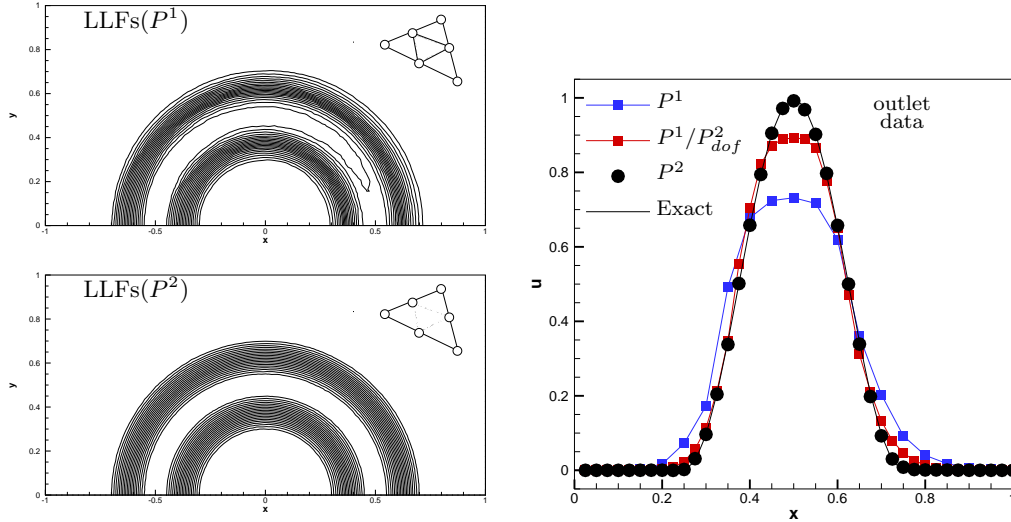
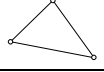
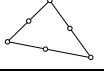



Figure 3.14: Rotation of a  $\cos^2$  profile :  $P^1$  and  $P^2$  LLFs results

We can now test the scheme defined by  $\phi_i^K = \phi_i^* + \phi_i^{SD_h}$  which we will denote by LLFs scheme, the s standing for *stabilized*. Figure 3.14 reports the results for the rotation of the  $\sin^2$  profile obtained with the  $P^1$  and  $P^2$  scheme : the results are now perfectly smooth, and, as the line plot on the right shows, the LLFs( $P^2$ ) over performs the LLFs( $P^1$ ) on the conformal  $P^1$  sub-triangulation (comparison with same degrees of freedom). The accuracy properties are verified by refining the grid, as shown on table 3.1. As visible on the table, and on the picture next to it, for a given number of degrees of freedom, the error is reduced of one order of magnitude going from  $P^k$  to  $P^{k+1}$ .

			
$h$	$\epsilon_{L^2}(P^1)$	$\epsilon_{L^2}(P^2)$	$\epsilon_{L^2}(P^3)$
1/25	0.50493E-02	<b>0.32612E-04</b>	<b>0.12071E-05</b>
1/50	0.14684E-02	0.48741E-05	<b>0.90642E-07</b>
1/75	0.74684E-03	0.13334E-05	0.16245E-07
1/100	<b>0.41019E-03</b>	<b>0.66019E-06</b>	0.53860E-08
	$\mathcal{O}_{L^2}^{ls} = 1.790$	$\mathcal{O}_{L^2}^{ls} = 2.848$	$\mathcal{O}_{L^2}^{ls} = 3.920$

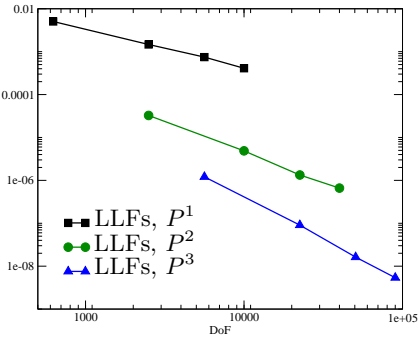


Table 3.1: Grid convergence and error *vs* DoF

The definition of the smoothness sensor is tested on the Burger’s test. The comparison of the  $P^1$  and  $P^2$  results is shown on figure 3.15. We can see that no oscillations are produced, and that the discontinuity is captured in one element, hence the  $P^2$  shock layer contains more nodes. This is a major difference with respect to the MU schemes of the previous section,

capable of capturing both shocks and shears within one *sub-element*.

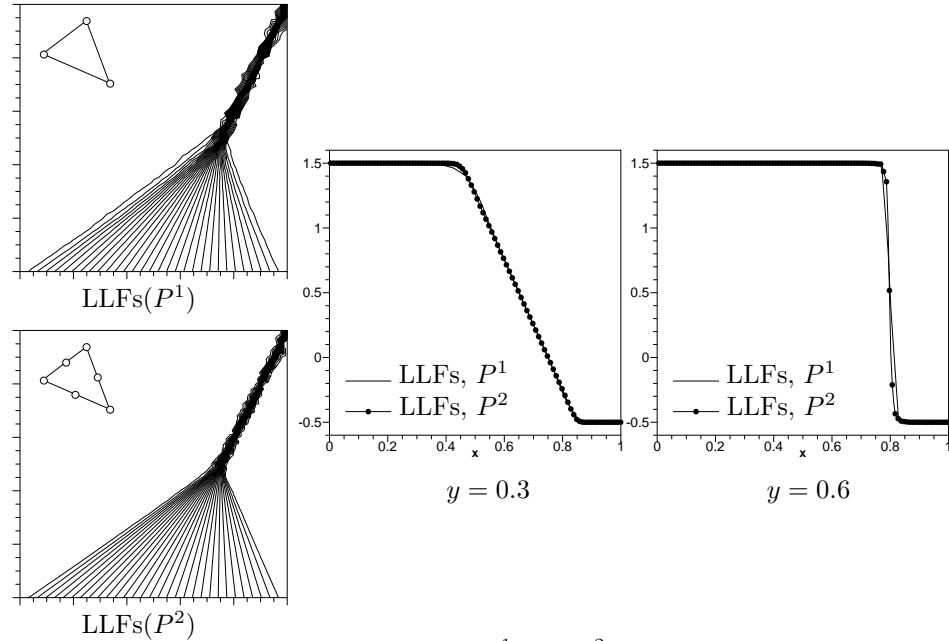


Figure 3.15: Burger's equation :  $P^1$  and  $P^2$  LLFs scheme results

The schemes can be formally generalized to the Euler equations of gas dynamics. This is done by using the limiter (3.4)-(3.6), evaluating the smoothness sensor on the density, and using a matrix formulation of the SD term. Details are given in [ALR11].

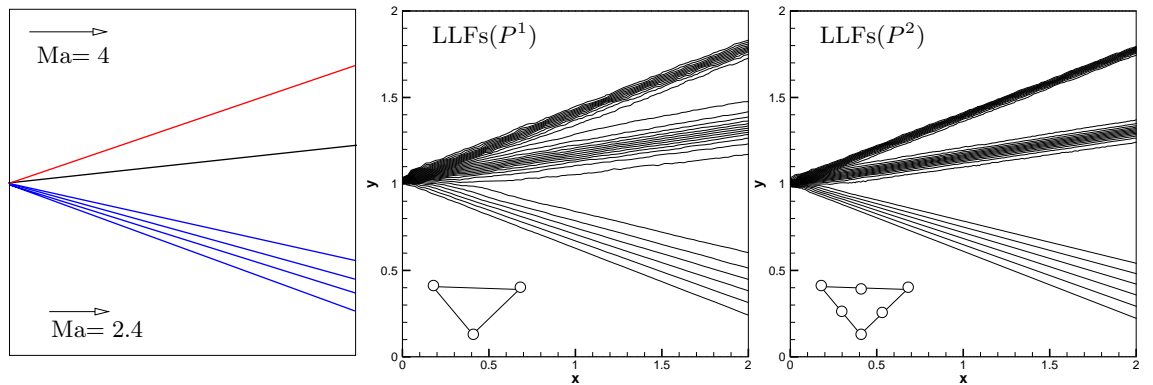


Figure 3.16: Supersonic jets interaction : problem definition and density contours

We present here a few examples to illustrate the behavior of the schemes. The first is the interaction of two supersonic jets, leading to the formation of an oblique shock, a contact and a supersonic expansion fan. A sketch of the problem and the results of the LLFs( $P^1$ ) and LLFs( $P^2$ ) schemes on an unstructured triangulation with the topology on the left on

figure 3.8 are reported on figures 3.16 and 3.17. The results show a clean capturing of all the discontinuities. Looking at the outlet data on figure 3.17, we can see that while the entropy profiles are nicely monotone, a small undershoot is visible in the Mach distribution. Perhaps a sign that the shock sensor should not only take into account the density but also some hydrodynamic quantity such as *e.g.* kinetic energy.

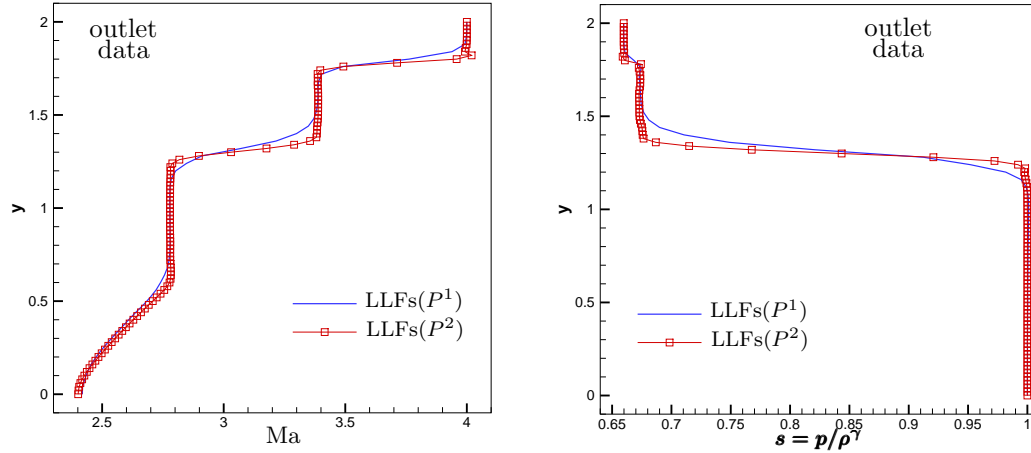


Figure 3.17: Supersonic jets interaction : outlet Mach number and entropy

We finally report two results taken from [ALR11] to confirm the accuracy of the schemes. The first involves the subsonic Mach 0.5 flow around a circular cylinder. The computation is run on a hybrid  $P^k - Q^k$  mesh shown on the top-left on figure 3.18. The computations have been run for  $k = 1$  and  $k = 2$ , using the conformally refined grid in the  $P^1 - Q^1$  case in order to compare results on the same number of degrees of freedom. Pressure and entropy contours are reported on figure 3.18. The third order results are clearly better than the second order ones (same contours are plotted) ; the pressure profile is more symmetric, and the entropy generation is much lower.

Lastly, on figure 3.19 we report the results on the Ringleb flow, an exact solution of the Euler equations initially proposed by F. Ringleb [213]. we refer to [262, ALR11] for details concerning the exact solution. Figure 3.19 shows the grid convergence analysis performed in [ALR11] confirming the expected accuracy for both triangles and quadrilaterals, and also showing the net advantage of the richer  $Q^k$  approximation over the  $P^k$ , as seen from the much smaller absolute value of the error.

We refer to [ALRT09, ALR11] for more details and results to [ABJR11] for the application to three dimensional aerodynamics simulations.

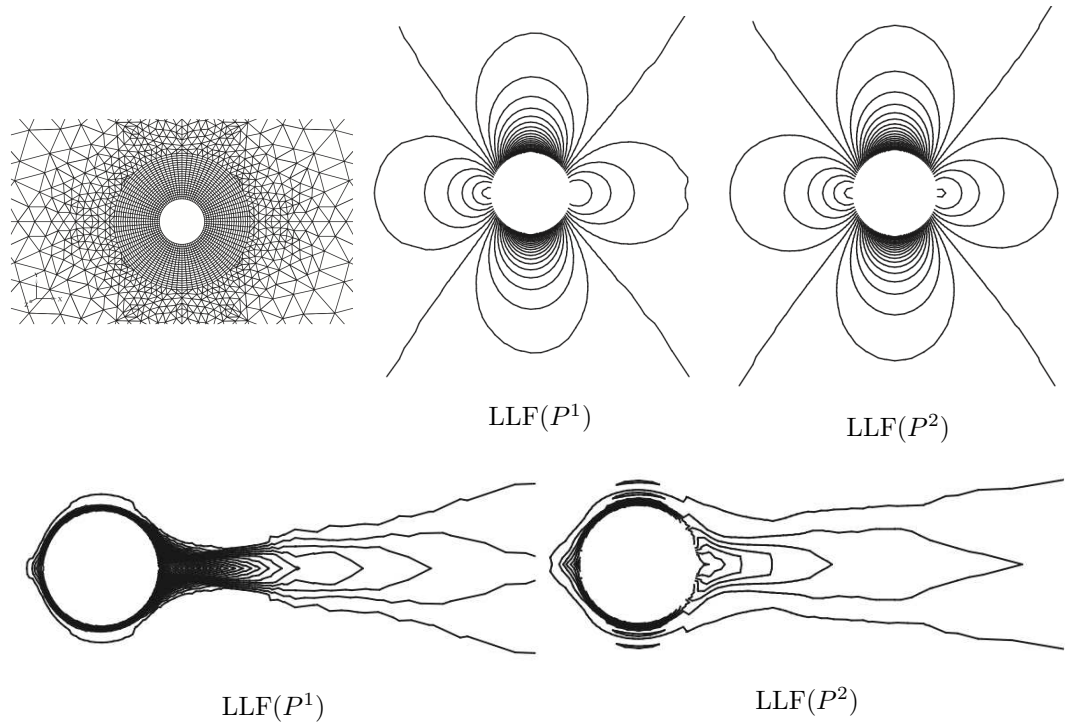


Figure 3.18: Subsonic flow around a circular cylinder :  $LLFs(P^1)$  and  $LLFs(P^2)$  results. Top row : grid and pressure contours. Bottom row : entropy variation

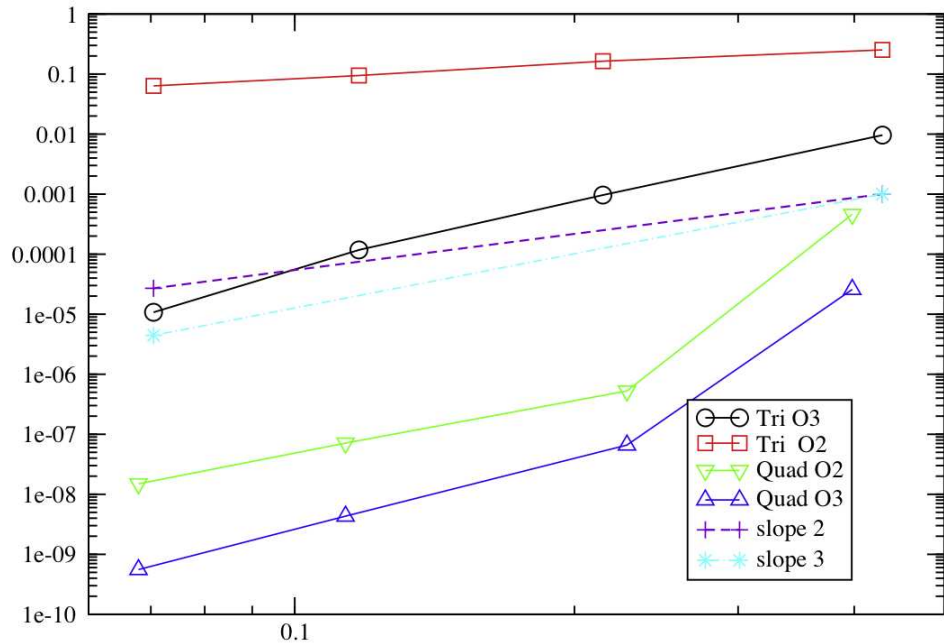


Figure 3.19: Ringleb flow : grid convergence analysis

# Chapter 4

## Contributions : high order schemes for time dependent conservation laws

### 4.1 Accuracy and time dependent conservation laws

In this chapter, we consider the approximation of time dependent solutions to

$$\partial_t u + \nabla \cdot \mathcal{F}(u) = 0 \quad \text{on} \quad \Omega \times [0, T_{\text{fin}}] \subset \mathbb{R}^d \times \mathbb{R}^+ \quad (4.1)$$

with all the hypotheses recalled in the beginning of section §2.1. As recalled in section §2.4, the discrete prototype (2.11) (or (2.16) including boundary conditions) is only first order accurate *in space*, with the exception of Taylor-Galerkin/Lax-Wendroff type schemes recast in a RD formalism [RD99, 145, CdSR<sup>+</sup>00, 224].

In this section, we will recall the general integral truncation error analysis for (4.1), after [RAD07], giving conditions on a scheme of the RD type for providing a higher order approximation of (4.1) in the time dependent case. These conditions are used to construct different version of the schemes for time dependent simulations.

The material discussed in this chapter is inspired by the work published in [RCD01, CRDP01, CRD03a, CRD03b, Ric05, DR07, HR09, HR10, HR11], [MRAD03, RCD04, RCD05, RA06, RB09b], and [RA10].

#### 4.1.1 RD prototype for time dependent solutions

We introduce the time discretized version of (4.1) by means of an  $r+1$ th order time integration scheme

$$\Gamma^{n+1}(u) = \sum_{i=0}^p \alpha_i \frac{\delta u^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \nabla \cdot \mathcal{F}^{n+1-j} \quad (4.2)$$

with  $\delta u^{n+1} = u^{n+1} - u^n$ ,  $\mathcal{F}^{n+1-j} = \mathcal{F}^{n+1-j}(u^{n+1-j})$ , and with the  $\alpha_i$  and  $\theta_j$  coefficients given by a time integration scheme of choice. In particular, we assume that the time stepping



scheme verifies the conservation identity

$$\sum_{n=0}^N \sum_{i=0}^p \alpha_i \delta u^{n+1-i} = u^N - u^0 = u(\text{T}_{\text{fin}}) - u_0 \quad (4.3)$$

We set on every  $K \in \Omega_h$

$$\Phi^K = \int_K \Gamma^{n+1}(u_h) = \int_K \left( \sum_{i=0}^p \alpha_i \frac{\delta u_h^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \nabla \cdot \mathcal{F}_h^{n+1-j} \right) \quad (4.4)$$

where  $u_h$  is some continuous discrete polynomial approximation of the type

$$u_h = \sum_{i \in \Omega_h} \varphi_i u_i$$

with  $\varphi_i$   $k$ th degree Lagrange (or other [13]) basis functions. Similarly,  $\mathcal{F}_h$  is a  $k+1$ th order accurate flux approximation. Similarly, on each boundary face  $f$  we set

$$\phi^f = \int_f \gamma^{n+1}(\vec{n}) = \int_f \sum_{j=0}^q \theta_j (G^* - \mathcal{F}_h)^{n+1-j} \cdot \vec{n} \quad (4.5)$$

with  $G^*$  a numerical flux consistent with the BCs.

We consider the scheme that computes  $u_h$  as the solution of

$$\sum_{K \in K_i} \Phi_i^K + \sum_{f \in F_i} \phi_i^f = 0 \quad (4.6)$$

where  $\forall K$  and  $\forall f$

$$\sum_{j \in K} \Phi_j^K = \Phi^K \quad \text{and} \quad \sum_{j \in f} \phi_j^f = \phi^f \quad (4.7)$$

**Remark 4.1.1** (Time stepping schemes). *The definition of  $\Gamma^{n+1}(u)$  is meant to accommodate not only most multi-step (including single step RK-type methods) scheme, but also space time and Galerkin (or Petrov-Galerkin) time discretizations. The last case easily fits in for example in the case of a direct time interpolation of the flux (not necessarily of the same polynomial degree of  $u$ ).*

#### 4.1.2 Accuracy analysis

We follow [RAD07, 13]. To simplify the notation we consider the scalar case, and we neglect the boundary conditions, which are supposed to be exactly satisfied (so that  $\gamma(\vec{n}) = 0$  everywhere). The system case is easily obtained by replacing absolute values by norms. For the inclusion of the boundary conditions, the reader can consult [ALR11, 13].

We will assume that the mesh and the time stepping strategy satisfy the regularity assumptions

$$C_0 \leq \sup_{K \in \Omega_h} \frac{h^2}{|K|} \leq C_1, \quad C'_0 \leq \frac{\Delta t}{h} \leq C'_1 \quad (4.8)$$

where we recall that  $\Delta t = \min_n(t^{n+1} - t^n)$ , with  $\Delta t^{n+1} = t^{n+1} - t^n$ .

Let now  $u$  be a regular  $u \in C^{l+1}$  exact classical solution of (4.1), with  $l \geq \max(r, k)$ , and such that

$$\sum_{i=0}^p \alpha_i \frac{\delta u^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \nabla \cdot \mathcal{F}^{n+1-j} = \partial_t u + \nabla \cdot \mathcal{F} + \mathcal{O}(\Delta t^{r+1}) \quad (4.9)$$

Consider now  $u_h^n$ , the  $k+1$ th order accurate approximation of  $u$  obtained with a continuous Lagrange interpolation of  $u$ .

Consider now  $\psi \in C_0^1(\Omega \times [0, T_{\text{fin}}])$ , a smooth test function with  $\psi|_{\partial\Omega} = 0$ . Let  $\psi_i^n$  be its nodal values  $\psi_i^n = \psi(\bar{x}_i, t^n)$ , and consider the  $k+1$ th order accurate space-time approximation  $\psi_h$ . It is also assumed that [63, 105] there exist constants  $C_0''$ ,  $C_1''$ ,  $C_2$  such that

$$\begin{aligned} \|\partial_t \psi_h\|_{L^\infty(\Omega_h)} &\leq C_0'', & \|\psi_h(t + \Delta t) - \psi_h(t)\|_{L^\infty(\Omega_h)} &\leq C_0'' \Delta t \\ \|\psi_h\|_{L^\infty(\Omega_h)} &\leq C_1'', & |\psi_i - \psi_j| &\leq \|\nabla \psi_h\|_{L^\infty(\Omega_h)} h \leq C_2 h \end{aligned} \quad (4.10)$$

We define the following truncation error for scheme (4.6)

$$\epsilon(u_h, \psi) := \sum_{n=0}^N \sum_{i \in \Omega_h} \Delta t^{n+1} \psi_i^{n+1} \sum_{K \in K_i} \Phi_i^K(u_h) = \sum_{n=0}^N \sum_{K \in \Omega_h} \sum_{i \in K} \int_{t^n}^{t^{n+1}} \psi_i^{n+1} \Phi_i^K(u_h) \quad (4.11)$$

We introduce the Galerkin splitting

$$\Phi_i^G = \int_K \varphi_i \Gamma^{n+1}$$

and note that

$$\sum_{j \in K} (\Phi_j^K - \Phi_j^G) = 0$$

This allows to recast the error as

$$\epsilon(u_h, \psi) = \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \left\{ \int_{\Omega_h} \psi_h^{n+1} \Gamma^{n+1}(u_h) + \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{i, j \in K} (\psi_i - \psi_j) (\Phi_i^K - \Phi_i^G) \right\} \quad (4.12)$$

Multiplying (4.9) by  $\psi_h$  and integrating over space and time we can get

$$\sum_{n=0}^N \int_{t^n}^{t^{n+1}} \int_{\Omega_h} \psi_h^{n+1} \Gamma^{n+1}(u) = \sum_{n=0}^N \Delta t \mathcal{O}(\Delta t^{r+1}) = \mathcal{O}(\Delta t^{r+1})$$

So the error can be estimated as follows

$$\begin{aligned} \epsilon(u_h, \psi) &= \text{I} + \text{II} + \text{III} + \mathcal{O}(\Delta t^{r+1}) \\ \text{I} &= \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \int_{\Omega_h} \psi_h^{n+1} \sum_{i=0}^p \alpha_i \frac{\delta(u_h - u)^{n+1-i}}{\Delta t} \\ \text{II} &= \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{j=0}^q \int_{\Omega_h} \psi_h^{n+1} \nabla \cdot (\mathcal{F}_h - \mathcal{F})^{n+1-j} \\ \text{III} &= \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{i, j \in K} (\psi_i - \psi_j) (\Phi_i^K - \Phi_i^G) \end{aligned}$$

Estimating each of the terms we get to the conditions of the cell and boundary splittings allowing to preserve the  $\mathcal{O}(\Delta t^{r+1})$  appearing on the right hand side. This is readily done by using the hypotheses on the regularity of  $u$  and standard interpolation results [63, 105]. In particular, using hypothesis (4.3) we rewrite term I as

$$\begin{aligned} \text{I} &= \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \int_{\Omega_h} \sum_{i=0}^p \alpha_i \frac{\delta(\psi_h u_h - \psi_h u)^{n+1-i}}{\Delta t} - \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \int_{\Omega_h} (u_h - u) \sum_{i=0}^p \alpha_i \frac{\delta(\psi_h)^{n+1-i}}{\Delta t} \\ &= \int_{\Omega_h} (\psi_h(u_h - u))(\mathbb{T}_{\text{fin}}) - \int_{\Omega_h} (\psi_h(u_h - u))_0 - \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \int_{\Omega_h} (u_h - u) \sum_{i=0}^p \alpha_i \frac{\delta(\psi_h)^{n+1-i}}{\Delta t} \end{aligned}$$

Using (4.10), and the regularity of  $u$ , we can now bound this term as

$$|\text{I}| = \mathcal{O}(h^{k+1}) + C \frac{\mathbb{T}_{\text{fin}}}{\Delta t} \Delta t \mathcal{O}(h^{k+1}) C_0'' \sup_{i=1,p} |\alpha_i| = \mathcal{O}(h^{k+1})$$

Term II is estimated noting that because of the compact support of  $\psi$  and due to the continuity of the approximation

$$\begin{aligned} \int_{\Omega_h} \psi_h^{n+1} \nabla \cdot (\mathcal{F}_h - \mathcal{F})^{n+1-j} &= \sum_{K \in \Omega_h} \int_K \psi_h^{n+1} \nabla \cdot (\mathcal{F}_h - \mathcal{F})^{n+1-j} \\ &= \sum_{K \in \Omega_h} \oint_{\partial K} \psi_h^{n+1} (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \vec{n} - \sum_{K \in \Omega_h} \int_K (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \nabla \psi_h^{n+1} \\ &= \oint_{\partial \Omega_h} \psi_h^{n+1} (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \vec{n} - \int_{\Omega_h} (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \nabla \psi_h^{n+1} \\ &= - \int_{\Omega_h} (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \nabla \psi_h^{n+1} \end{aligned}$$

So that using (4.10), the regularity of  $u$  and the accuracy of the approximation :

$$|\text{II}| \leq \left| \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{j=1}^q \int_{\Omega_h} (\mathcal{F}_h - \mathcal{F})^{n+1-j} \cdot \nabla \psi_h^{n+1} \right| \leq C \frac{\mathbb{T}_{\text{fin}}}{\Delta t} \Delta t \mathcal{O}(h^{k+1}) \sup_{j=1,q} |\theta_j| C_2 = \mathcal{O}(h^{k+1})$$

To conclude the analysis we remark that the Galerkin splitting can be re-manipulated as we have already done for the global integrals. In particular, we can write

$$\Phi_i^G(u_h) = \Phi_i^G(u_h - u) = \Phi_i^G(u_h - u) + \mathcal{O}(h^2) \mathcal{O}(\Delta t^{r+1})$$

and recast the remaining terms as

$$\text{III} = \mathcal{O}(h^2) \mathcal{O}(\Delta t^{r+1}) + \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \frac{1}{C_K} \sum_{K \in \Omega_h} \sum_{i,j \in K} (\psi_i - \psi_j) (\Phi_i^K - \Phi_i^G(u_h - u))$$

where

$$\Phi_i^G(u_h - u) = \sum_{i=0}^p \alpha_i \int_K \varphi_i \frac{\delta(u_h - u)^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \int_K \varphi_i \nabla \cdot (\mathcal{F}_h - \mathcal{F})^{n+1-j}$$

Using the accuracy of the approximation and the smoothness of  $u$  we can write

$$|\Phi_i^G(u_h - u)| \leq \sup_{i=1,p} |\alpha_i| C \frac{h^2}{\Delta t} \mathcal{O}(h^{k+1}) + \sup_{j=0,q} |\theta_j| C h^2 \mathcal{O}(h^k) = \mathcal{O}(h^{k+3})\mathcal{O}(\Delta t^{-1}) + \mathcal{O}(h^{k+3})$$

Using again (4.10)

$$|\text{III}| \leq \mathcal{O}(h^2)\mathcal{O}(\Delta t^{r+1}) + C \frac{T_{\text{fin}}}{\Delta t} \frac{\Delta t}{C_K} \frac{|\Omega|}{h^2} C_2 h \left( \mathcal{O}(h^{k+3})\mathcal{O}(\Delta t^{-1}) + \mathcal{O}(h^{k+3}) + \sup_{K \in \Omega_h} \sup_{i \in K} |\Phi_i^K| \right)$$

So that collecting all the contributions, the error can be bounded as

$$|\epsilon(u_h, \psi)| \leq \mathcal{O}(\Delta t^{r+1}) + \underbrace{\mathcal{O}(h^{k+1})}_{\text{from I and II}} + \underbrace{\mathcal{O}(h^{k+2})\mathcal{O}(\Delta t^{-1}) + \mathcal{O}(h^{k+2})}_{\text{from III}} + \underbrace{\mathcal{O}(h^{-1}) \sup_{K \in \Omega_h} \sup_{i \in K} |\Phi_i^K|}_{\text{from III}} \quad (4.13)$$

and finally, because of the regularity assumptions (4.8)

$$|\epsilon(u_h, \psi)| \leq \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{-1}) \sup_{K \in \Omega_h} \sup_{i \in K} |\Phi_i^K| \quad (4.14)$$

with  $p = \min(k, r)$ . This estimate leads to the result that follows.

**Proposition 4.1.2** (Accuracy of RD, unsteady case). *Under assumption (4.8) on the time stepping, given a  $k + 1$ th order continuous polynomial approximation of the unknown ad of the fluxes, and a  $r + 1$ th order accurate time integration scheme, scheme (4.6) verifies the truncation error estimate*

$$|\epsilon(u_h, \psi)| \leq \mathcal{O}(h^{p+1}), \quad p = \min(k, r)$$

provided that

$$\sup_{K \in \Omega_h} \sup_{i \in K} |\Phi_i^K(u_h)| = \mathcal{O}(h^{p+2}) \quad (4.15)$$

whenever  $u_h$  is the interpolant of a smooth exact solution. In this case we say that the scheme is  $p + 1$ th order accurate.

Moreover we have the following estimate.

**Lemma 4.1.3** (Consistency estimate, time dependent case). *Under the hypotheses of proposition 4.1.2 the following consistency estimates hold.*

$$\Gamma^{n+1}(u_h) = \mathcal{O}(h^k), \quad \Phi^K(u_h) = \mathcal{O}(h^{k+2}) \quad (4.16)$$

*Proof.* The proof is easily obtained by considering that due to (4.9)

$$\Gamma^{n+1}(u_h) = \mathcal{O}(\Delta t^{r+1}) + \Gamma^{n+1}(u_h) - \Gamma^{n+1}(u)$$

By its definition, and under the hypotheses made, one easily checks that

$$\Gamma^{n+1}(u_h) - \Gamma^{n+1}(u) = \mathcal{O}(h^{k+1})\mathcal{O}(\Delta t^{-1}) + \mathcal{O}(h^k) = \mathcal{O}(h^k)$$

□

As a consequence we have the following corollary.

**Corollary 4.1.4** (High order residual schemes). *Under the hypotheses of proposition 4.1.2, a sufficient condition for a scheme of the form (4.6) to be  $p + 1$ th order accurate if there exist a test function  $\omega_i$  uniformly bounded w.r.t.  $h$ ,  $u_h$ ,  $\Gamma^{n+1}(u_h)$ , and w.r.t the data of the problem, such that*

$$\Phi_i^K(u_h) = \int_K \omega_i \Gamma^{n+1}(u_h) \quad (4.17)$$

**Remark 4.1.5** (Boundary conditions). *The analysis can be extended to include boundary conditions, with some additional effort on the description of the geometrical discretization of the boundary  $\partial\Omega_h$ , on which the numerical BCs are imposed, and with some regularity hypotheses on  $\partial\Omega$  on which the BCs are imposed for the exact solution. When the difference between  $\partial\Omega_h$  and  $\partial\Omega$  is neglected, the analysis extends quite trivially, and both the consistency estimates and the accuracy conditions can be derived for the face residuals  $\phi_i^f$ . The final result is that a sufficient condition for a scheme of the form (4.6) to be  $p + 1$ th order accurate if there exist test functions  $\omega_i$  and  $\omega_i^f$  uniformly bounded w.r.t.  $h$ ,  $u_h$ ,  $\Gamma^{n+1}(u_h)$ ,  $\gamma^{n+1}(\vec{n})$ , and w.r.t the data of the problem, such that [13]*

$$\Phi_i^K(u_h) = \int_K \omega_i \Gamma^{n+1}(u_h) \quad \Phi_i^K(u_h) = \int_f \omega_i^f \gamma^{n+1}(\vec{n}) \quad (4.18)$$

### 4.1.3 Scheme zoology

The accuracy analysis shows that a sound way of reproducing the accuracy preserving property in the time dependent case is to construct the discretization starting from some sort of variational principle that couples all the terms in the equation, including the time derivative.

With no exceptions, this leads to the necessity of inverting a mass matrix. In particular, upwind methods will introduce a degree of upwinding of the time derivative which can hardly be removed. Unless one resorts to some other technique (*e.g.* Taylor-Galerkin [145, RD99, 224]), three possible approaches can be investigated :

**Space-time schemes** If problem (4.1) is thought as a sequence of steady problems in space-time within a set of temporal slabs  $[t^n, t^{n+1}]$ , plenty of possibilities are available, starting from the use of standard fluctuation splitting on space-time simplicia, to the (more intelligent) use of schemes on prismatic extruded space-time meshes. These ideas are explored in [RCD01, CRDP01, CRD03a, CRD03b, Ric05, DR07, HR09, HR10, HR11] in collaboration with the group at the von Karman Institute, and, more recently, with M. Hubbard of the School of computing at Leeds University, and A. Larat of the École Central de Paris. Space time schemes are described in section §4.2 ;

**Implicit schemes with mass matrix** The alternative is to find smart constructions of mass matrices/test functions to couple with some temporal integration strategy. This was the initial idea behind the work of Maerz and Ferrante at the von Karman Institute [174, 107] and of Caraeni at Lund University [54]. This approach, recalled already in section §2.4 has been pursued first in [9, 180], and later in [MRAD03, RCD04, RCD05, RA06, RB09b], in collaboration with my colleagues at INRIA and at the von Karman Institute. These developments are recalled in section §4.3 ;

**Genuinely explicit schemes** The last option is to keep one's hopes of building a genuinely explicit scheme that still has a residual character and that incorporates ideas such as Multidimensional Upwinding. This has to be done keeping into account corollary 4.1.4. This has been achieved in [RA10], and is the object of section §4.4.

## 4.2 High order space-time formulations

### 4.2.1 Space-time schemes on triangles and tets

The simplest (and craziest) idea one can have is to re-cycle known fluctuation splitting schemes using linear space-time elements. At a first glance it might seem like an easy thing to do. Actually, due to the underlying continuous approximation, it turns out to be a tricky business, unless one wants to solve in one shot the entire time dependent problem (4.1) on a  $d + 1$ -dimensional grid.

The main issue that has to be dealt with is how to make sure that we can march in time and that not too much information travels back toward the past. The other question is how to make efficient a procedure that might be highly implicit, and require a (relatively) high computational time per time-step, compared to a purely explicit single or multi step scheme. A similar development was done in the DG framework by R.B.Lowrie in his PhD [173]. As shown in the reference, one of the keys to the solution of these issues is a smart use of upwinding.

In the fluctuation splitting context, answers to these questions were given in [CRDP01, CRD03b, CRD03a] (see also the PhD of A.Csik [79, 80]). The idea in these papers is that the discretization procedure should be exactly the same as the one used for the solution of steady problems. The difference is that the fluctuation should be computed as

$$\Phi^{K_t} = \int_{K_t} (\partial_t u_h + \vec{a} \cdot \nabla u_h)$$

where  $K_t$  is a space time simplex. If a linear variation of the solution on the space-time element is assumed, then, exactly as in the steady case, the element residual can be expressed as

$$\Phi^{K_t} = \sum_{K_t} k_j u_j$$

where now the inflow parameters are easily shown to be given by

$$k_i = \frac{1}{d+1} (\vec{a}_{K_t} \cdot \vec{n}_i^S + n_i^t), \quad (4.19)$$

where  $d$  denotes the number of space dimensions,  $\vec{a}_{K_t}$  is a local average of the flux Jacobian (2.2) over  $K_t$ , and where the superscripts  $S$  and  $t$  denote the spatial and temporal components of the element normals.

The key is to use Multidimensional Upwinding (cf. section §2.2.4) as a means of guaranteeing that no information travels toward the past. To do this, one makes sure that  $k_i \leq 0$  whenever node  $i$  is in the past. Let us consider for example the one dimensional problem

$$\partial_t u + a \partial_x u = 0$$

Consider a time slab between two constant time levels. One easily realizes that two types of elements can exist in a 2D space-time mesh filling the slab. As shown on the left picture on figure 4.1, type 1 elements have one node in the future and two in the past, while type 2 elements have only one node in the past.

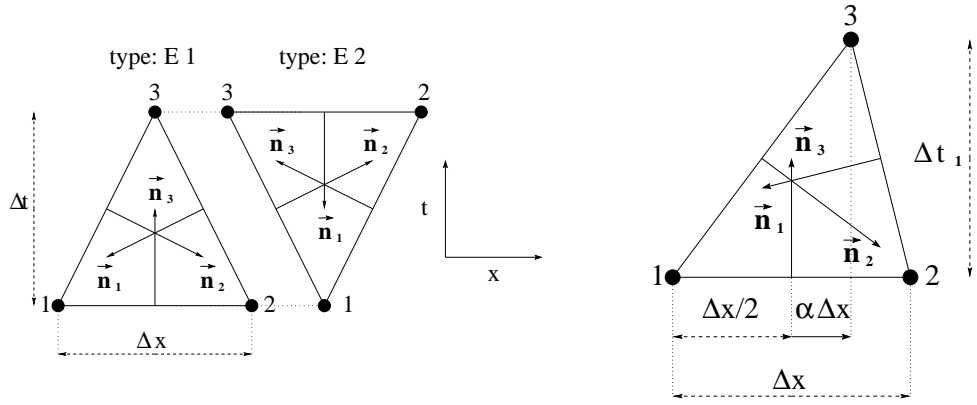


Figure 4.1: Geometry for space-time schemes on linear triangles

Type 2 elements pose no problem. In these elements  $\vec{n}_i^S = 0$  for the only node in the past (node 1 in the left picture on figure 4.1), and  $n_i^t = -\Delta x$  so that  $k_i < 0$  always. As a consequence, any MU scheme will not update this node, thus allowing to march in the correct time direction.

Type 1 elements deserve more attention. Consider the case of a general type 1 element shown in the right picture on figure 4.1. Let the parameter  $\alpha$  define the spatial location of the node at the intermediate time level. For  $|\alpha| > \frac{1}{2}$ , one obtains an obtuse triangle (*i.e.* the projection of node 3 falls outside the edge located at level  $n$ ), while  $\alpha = 0$  corresponds to the symmetric case. The space-time face normals are easily computed :

$$\vec{n}_1 = (-\Delta t, -(\frac{1}{2} - \alpha)\Delta x), \quad (4.20)$$

$$\vec{n}_2 = (\Delta t, -(\frac{1}{2} + \alpha)\Delta x), \quad (4.21)$$

$$\vec{n}_3 = (0, \Delta x). \quad (4.22)$$

The decoupling condition for nodes 1 and 2 not to receive any information is again  $k_1, k_2 \leq 0$ , which, using (4.19) becomes

$$\begin{aligned} k_1 &= -\frac{a\Delta t}{2} - \left(\frac{1}{2} - \alpha\right)\frac{\Delta x}{2} \leq 0 \\ k_2 &= \frac{a\Delta t}{2} - \left(\frac{1}{2} + \alpha\right)\frac{\Delta x}{2} \leq 0 \end{aligned} \quad (4.23)$$

We see right away that, whatever the sign of  $a$ , one of the two conditions will impose a time step limitation. Moreover, no positive solution for  $\Delta t$  exists for  $|\alpha| \geq \frac{1}{2}$ , excluding obtuse and rectangle triangles of type E1 in the first layer, thus ruling out the upper mesh topology in the left picture on figure 4.2.

For  $|\alpha| < \frac{1}{2}$  the time step limitation on the first layer is

$$\nu = \frac{\Delta t |a|}{\Delta x} \leq \frac{1}{2} - |\alpha|. \quad (4.24)$$

making  $\alpha = 0$  the best choice. This choice corresponds to the lower mesh topology in the left picture on figure 4.2.

Preliminary results based on this approach were shown in [80]. Using Multidimensional upwind schemes such as the LDA or the PSI scheme, and provided that (4.24) is met, second order is obtained on time dependent problems by solving

$$\sum_{K_i | K \in K_t \cap K \in K_i} \beta_i^{K_t} \phi^{K_t} = 0$$

Time marching is obtained by solving every time on a space-time mesh of with the lower mesh topology in the left picture on figure 4.2, on the temporal slab  $[t^n, t^{n+1}]$ . The major drawback of this approach is that, even being highly implicit in nature, it is constrained by a explicit CFL =1/2 time step restriction.

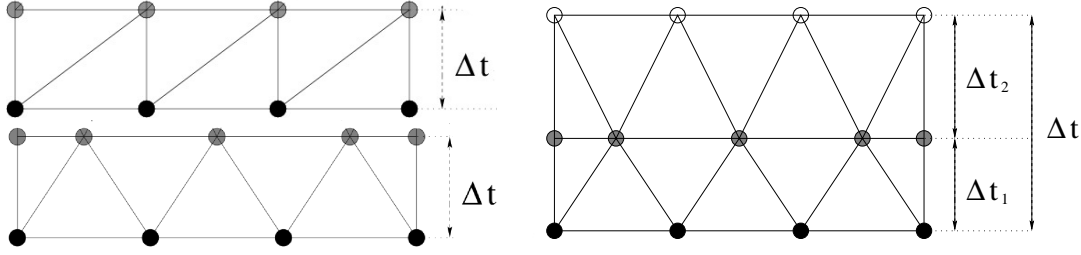


Figure 4.2: Space-time triangular mesh in 1D

The solution to this problem, proposed in [CRDP01, 80, CRD03b, CRD03a], is to couple two time slabs in each computation, and allow a stronger coupling between past and future nodes in the second slab. In one space dimension, this is obtained by employing the space time mesh shown on the right on figure 4.2.

The global time step allowed is now

$$\Delta t = \Delta t_1 + \Delta t_2 = (1 + Q)\Delta t_1 \quad (4.25)$$

with  $\Delta t_1$  respecting (4.24). The effective CFL number is now

$$\text{CFL} = (1 + Q)\nu, \quad \nu \leq \frac{1}{2}$$

where  $Q$  can be arbitrarily large. Condition (4.24) has been baptized the *past shield condition* in [CRDP01, 80].



The construction has been generalized to 2 space dimensions, employing the space-time mesh composed of three basic types of tetrahedrons : the blue, yellow, and red tetrahedra reported on the right on figure 4.3. A first layer of elements is created starting from the 2D mesh, and then mirrored to obtain the mesh of the second temporal slab.

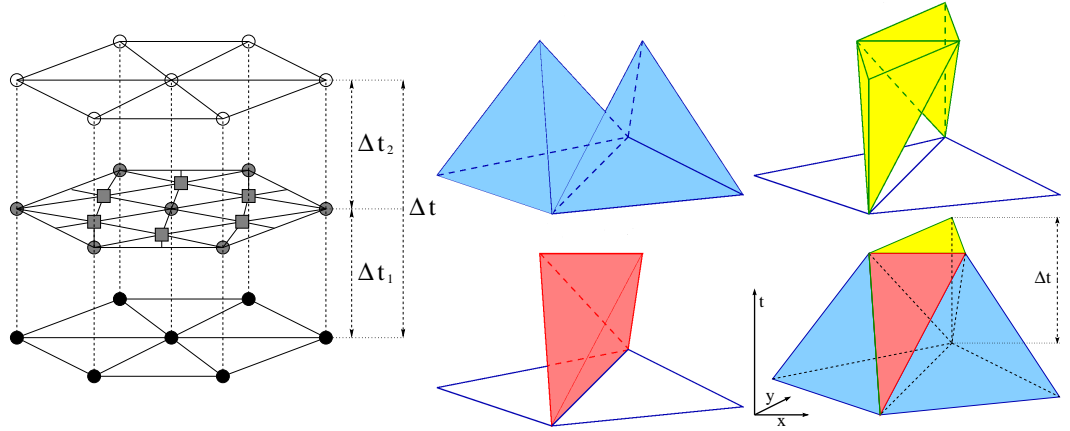


Figure 4.3: Space-time mesh in 2D. Left : overview of mirrored mesh. Right : basic tetrahedra composing the grid

For a given triangulation of the spatial domain  $\Omega_h$ , the space-time grid constructed in this way satisfies the past shield condition provided that  $\forall K \in \Omega_h$  [CRDP01, CRD03b]

$$\Delta t_1 \leq \min_{\substack{j \in K \\ k_j^+ > 0}} \frac{|n_j^t|}{k_j^+}$$

where now  $k_j$  denotes the standard inflow parameter (cf equation (2.27) and section §2.2.4)

$$k_j = \frac{\vec{a} \cdot \vec{n}_j}{2}$$

As before, the global time step is

$$\Delta t = \Delta t_1 + \Delta t_2 = (1 + Q)\Delta t_1$$

where  $Q$  is taken arbitrarily large.

Given the two-dimensional spatial grid, the three dimensional space time mesh is generated (once and for all in the beginning of the computation). The discrete solution of (4.1) is computed as the steady limit of a pseudo time stepping procedure [RCD01, CRDP01, CRD03b]

$$u_i^{k+1} = u_i - \omega_i \sum_{K_t | i \in K_t} \beta_i^{K_t} \Phi^{K_t} \quad (4.26)$$

For scalar problems, the results obtained show genuinely second order of accuracy on smooth solutions [CRDP01, CRD03b]. The extension to systems, can be done either by means of a Roe linearization, that trivially extends to this framework [CRDP01, CRD03b], or using the conservative approach of section §3.1 [RCD01]. An example of such an application is

reported on figure 4.4. The figure reports the solutions obtained with the blended LDA-N scheme (cf. equation (2.46) and section §2.2.6) on the Mach 3 wind tunnel with a forward facing step test case, initially proposed in [274]. As in [71], the grid size is equal to  $1/80$  (the finest mesh used in [274]) however, it is refined to  $10^{-3}$  in correspondence of the corner singularity, as shown on the top-left picture on figure 4.4. The capability for arbitrary time step has been used by choosing  $Q$  such that the global time step correspond to a CFL=1 computation in the non-refined region :

$$Q = 2 \frac{h_{\max}}{h_{\min}} - 1$$

The density contours of figure 4.4 (taken from [RCD01]) show a nice and clean capturing of all the flow features. More results, including the application to a two-phase flow model, can be found in [RCD01, CRDP01, CRD03b, CRD03a].

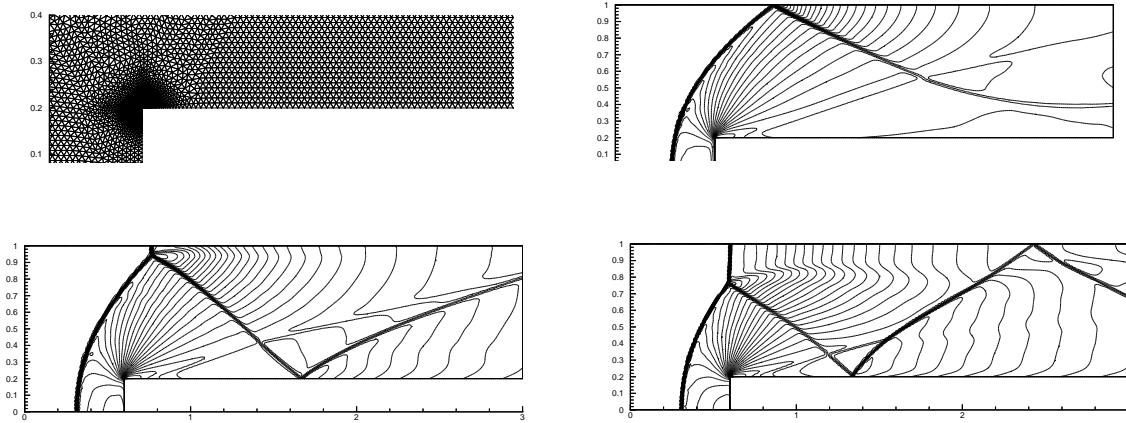


Figure 4.4: Mach 3 wind tunnel with a forward facing step. Blended LDA-N scheme on  $P^1$  space-time tetrahedra

The interest of this approach is that, once the details on the space-time meshing are worked out, it is relatively straightforward to code known schemes to solve time dependent problems. Even the extension to the laminar Navier-Stokes does not require much effort, as we have shown in [DRD02, DRD05].

The problem is of course that three dimensional computations require 4D meshes, and that the method is inherently very expensive. However, the lesson learned is that

- Space-time schemes based on formal extension of known RD schemes do yield high order of accuracy (see [RAD03] for more than second order) ;
- Multidimensional upwinding can be used in space time to design schemes that have a genuinely time marching character ;
- Unconditionally large time steps can be afforded at the cost of adding more variables ;
- Once an unconditionally stable time stepping space time scheme is obtained, all the properties of the scheme are retained for any value of the time step. This opens the door to the design of unconditionally positive and high order schemes.

### 4.2.2 Space-time schemes on extruded prisms

An improved space time formulation of RD is presented in [CRD03b, RCD04, RCD05, Ric05, DR07]. The idea is to use prismatic space time elements instead of tetrahedra. The missing piece, not allowing the use of this type of elements before, was the conservative formulation developed in [CRD02].

As done before, we define space-time cells. In this case, a space-time element is simply defined as  $K_t = K \times [t^n, t^{n+1}]$ . Over  $K_t$ ,  $u_h$  is represented using a bi-linear continuous approximation, obtained as the tensorial product of the space approximation times linear variation in time.

For scalar advection and in 2d, the idea is to rewrite the element residual as

$$\begin{aligned} \Phi^{K_t} &= \int_{K_t} (\partial_t u_h + \vec{a} \cdot \nabla u_h) = \sum_{j \in K} \left( \frac{\Delta t k_j}{2} + \frac{|K|}{3} \right) u_j^{n+1} + \sum_{j \in K} \left( \frac{\Delta t k_j}{2} - \frac{|K|}{3} \right) u_j^n \\ &= \sum_{j \in K} \bar{k}_j u_j^{n+1} + \sum_{j \in K} \hat{k}_j u_j^n \end{aligned} \quad (4.27)$$

Introducing the *space-time flux*  $(\vec{a}u, u) \in \mathbb{R}^2 \times \mathbb{R}$ , we can show that the  $\bar{k}_j$  and  $\hat{k}_j$  parameters, implicitly defined by (4.27), are the projection of the *space-time flux Jacobian*  $(\vec{a}, 1)$  along directions determined by the geometry of the prism  $K_t$ . To do this, we consider the shell  $\mathcal{S}_K$  formed by joining the gravity centers of  $K$  at times  $t^n$  and  $t^{n+1}$  with the nodes of the element at time  $t^{n+1/2} = t^n + (t^{n+1} - t^n)/2$  (left on figure 4.5). We can associate to each node of the prism the face of  $\mathcal{S}_K$  opposite to it, as illustrated on the right on figure 4.5 for node 1. With reference to this last picture, we introduce the space-time vectors  $\bar{n}_1$  and  $\hat{n}_1$ , normal to the faces of  $\mathcal{S}_K$  opposite to node 1, pointing inward with respect to the shell, and scaled by the area of the faces.

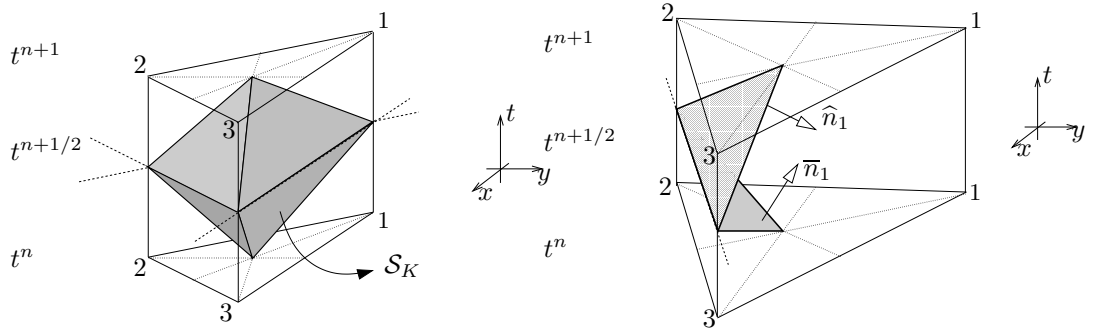


Figure 4.5: Closed shell  $\mathcal{S}_K$  (left), and space-time directions  $\bar{n}_1$  and  $\hat{n}_1$  (right)

Simple geometry shows that

$$\bar{k}_1 = \bar{n}_1 \cdot (\vec{a}, 1) \quad \text{and} \quad \hat{k}_1 = \hat{n}_1 \cdot (\vec{a}, 1)$$

Since  $(\vec{a}, 1)$  is the direction of a characteristic line cutting through the prism, we deduce that  $\bar{k}_1$  and  $\hat{k}_1$  are the projections of the direction of the characteristic onto  $\bar{n}_1$  and  $\hat{n}_1$ .

The idea is, as done on linear space-time elements, to use multidimensional upwind schemes to allow the propagation of the information in the correct time direction, ultimately allowing the construction of a time marching procedure. (see [DR07, Ric05] for more details)

on the geometry of the schemes). In particular, we look for high order schemes computing the nodal values of  $u_h$  as the solution of (BCs are not included) :

$$\begin{aligned} \sum_{K \in K_i} \bar{\beta}_i^{K_t} \Phi^{K_t} &= 0 \quad \text{for node } i \text{ at } t^{n+1} \\ \sum_{K \in K_i} \hat{\beta}_i^{K_t} \Phi^{K_t} &= 0 \quad \text{for node } i \text{ at } t^n \end{aligned} \quad (4.28)$$

where  $\forall K \in \Omega_h$

$$\sum_{j \in K} (\bar{\beta}_i^{K_t} + \hat{\beta}_i^{K_t}) = 1$$

We then define :

**Definition 4.2.1** (Space-time-MU scheme). *A space-time scheme is space-time Multidimensional Upwind (stMU) if in the prism  $K_t$*

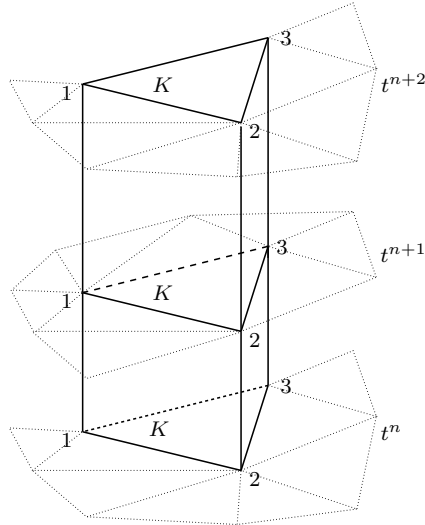
$$\begin{aligned} \bar{k}_j \leq 0 &\implies \bar{\beta}_i^{K_t} = 0 \\ \hat{k}_j \leq 0 &\implies \hat{\beta}_i^{K_t} = 0 \end{aligned}$$

And so we can generalize the past shield condition as follows.

**Proposition 4.2.2** (Space-time MU schemes and time-marching). *A stMU scheme defines a time-marching procedure if  $\hat{k}_j \leq 0, \forall j \in K$ , or equivalently*

$$\Delta t = t^{n+1} - t^n \leq \min_{K \in \Omega_h} \min_{j \in K} \frac{2|K|}{3k_j^+}, \quad \forall n \quad (4.29)$$

If condition (4.29) is verified, and a stMU scheme is used, we can remove the second equation in (4.28). Examples of stMU schemes are given in [CRD03b, DR07]. These are a straightforward generalization of the MU N, LDA, PSI (or limited N) and Blended schemes seen in sections §2.2.5 and §2.2.6.



Even if time marching is guaranteed by the respect of (4.29), scheme (4.28) still defines a highly implicit procedure for which a type time step limitation is a flaw. To overcome this limitations, as done in [RCD01, CRDP01], a second layer of prismatic cells can be added, as shown on the figure in the left. By allowing a coupling between the values of the solution at time  $n+1$  and  $n+2$ , arbitrarily large time steps can be taken. In particular, the global time step is

$$\Delta t = t^{n+2} - t^n = \Delta t_2 + \Delta t_1 = (1 + Q)\Delta t_1$$

with  $\Delta t_1 = t^{n+1} - t^n$  respecting (4.29), and  $Q$  arbitrary.

These schemes have been extensively tested in [CRD03b, Ric05] where it has been found that second order of accuracy is obtained with the stMU LDA scheme, and almost second order with the limited N and Blended LDA-N schemes (cf. sections §2.2.5 and §2.2.6). Using the conservative formulation of [CRD02] (see section §3.1), the schemes have been applied to the Euler equations in [Ric05, RCD05, DR07], to the shallow water equations in [RAD07], and to a compressible two-phase model in [RCD05, Ric05].

In particular, as an example, we consider the two phase model given by (4.1) with

$$u = \begin{bmatrix} \alpha_g \rho_g \\ \alpha_l \rho_l \\ \rho \vec{u} \end{bmatrix}, \quad \mathcal{F}(u) = \begin{bmatrix} \alpha_g \rho_g \vec{u} \\ \alpha_l \rho_l \vec{u} \\ \rho \vec{u} \otimes \vec{u} + p \mathbf{I} \end{bmatrix} \quad (4.30)$$

where  $\alpha_g$  and  $\alpha_l$  are the gas and liquid *volume fractions*,  $\rho_g$  and  $\rho_l$  are gas and liquid densities,  $\vec{u} = (u, v)$  is the local flow speed,  $\rho$  is the mixture density

$$\rho = \alpha_g \rho_g + \alpha_l \rho_l. \quad (4.31)$$

and  $p$  is the pressure. The model is closed by the relation

$$\alpha_g + \alpha_l = 1. \quad (4.32)$$

and by the EOS relating the densities to the pressure. In the following we will denote by  $\alpha$  the gas volume fraction, assuming implicitly that  $\alpha_l$  is obtained from (4.32). We will also refer to  $\alpha$  as to the *void* fraction. Concerning the EOS, we have used as in [200] the following relations representative of air and water (S.I. units are used):

$$p = \Gamma_g \left( \frac{\rho_g}{\rho_{g0}} \right)^{\gamma_g}, \quad p = \Gamma_l \left[ \left( \frac{\rho_l}{\rho_{l0}} \right)^{\gamma_l} - 1 \right] + p_{l0} \quad (4.33)$$

with  $\Gamma_g = 10^5$ ,  $\rho_{g0} = 1$ ,  $\gamma_g = 1.4$ , and  $\Gamma_l = 3.31 \times 10^8$ ,  $\rho_{l0} = 1000$ ,  $\gamma_l = 7.15$ , and  $p_{l0} = 10^5$ .

This system of equations constitutes a fairly simple model of homogeneous air-water two-phase flow. However, it has some appealing features for the purpose of testing our schemes. The first is precisely its simplicity, the second the fact that it is fully hyperbolic and its complete eigenstructure can be easily analytically derived. Most importantly, one can compute exact steady and unsteady Rankine-Hugoniot relations against which to test the schemes. In particular, with reference to the 1D shock depicted on the left on figure 4.6, on the right picture in the same figure we plot the pressure, void fraction and  $x$ -velocity ratios as functions of the Mach number

$$M_R = \frac{u_R}{\sqrt{p_R/\rho_R}} \quad (4.34)$$

As a consequence of the higher compressibility of the gas, the increase of pressure across shocks leads to a reduction of the gas volume fraction. Moving shocks are characterized in a similar way, by introducing the shock Mach number

$$M_S = \frac{u_S}{\sqrt{p_R/\rho_R}},$$

with  $u_S$  the velocity of the shock.

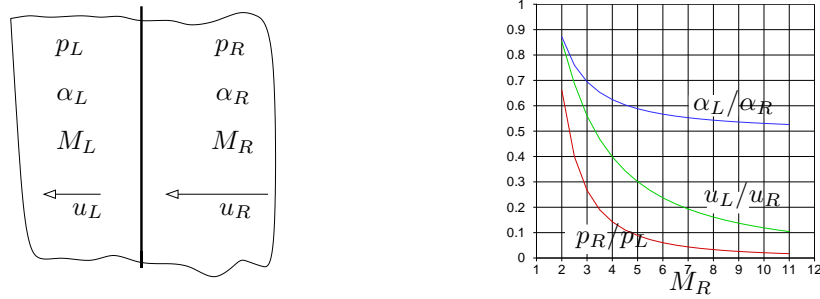


Figure 4.6: Jump conditions for the two-phase model. Flow from right to left.

Note however that the relation between the pressure and the conserved mass and momentum fluxes is so complex that a conservative linearization can hardly be derived. In particular, because of the nonlinearity of the equations of state, pressure and volume fractions cannot be computed in closed form from the conserved variables. Instead, combining the equations of state and relation (4.32), a nonlinear equation for the pressure is obtained which can be solved in a few Newton iterations (see [200] for more). In conclusion, even being so simple, this model has all the features of systems of conservation laws with complex thermodynamics.

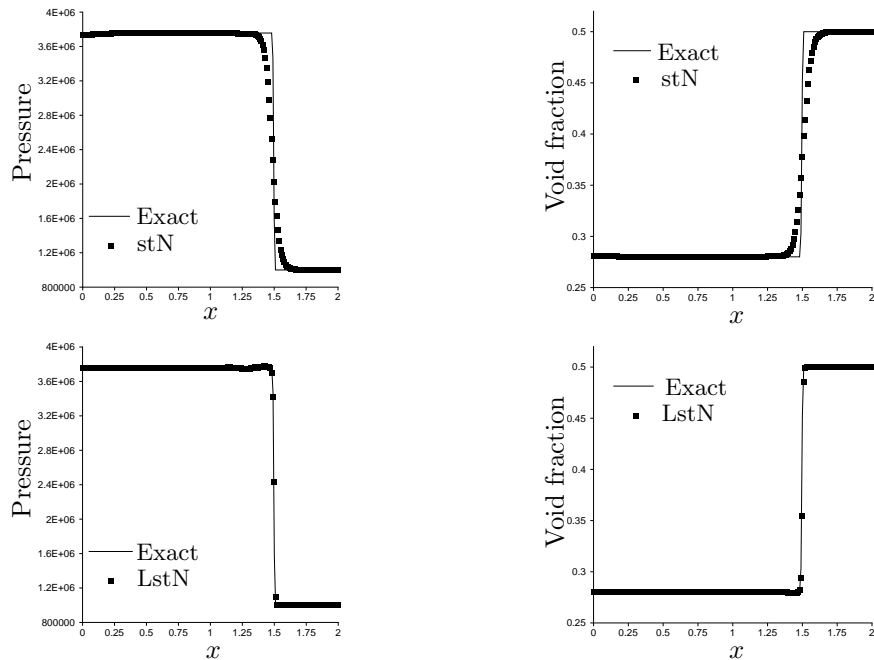


Figure 4.7: Two-phase  $M_S = 3$  shock. Pressure (left) and void fraction (right) along the line  $y = 0.05$ . Solutions of the stN (top, denoted as N2) and LstN (bottom) schemes

We present some results obtained with the single layer formulation, computing space-time element fluctuations as

$$\Phi^{K_t} = \int_{t^n}^{t^{n+1}} \int_K (\partial_t u_h + \nabla \cdot \mathcal{F}_h(u_h)) \approx \sum_{j \in K} \frac{K}{3} (u_j^{n+1} - u_j^n) + \Delta t \oint_{\partial K} \mathcal{F}(u_h^{n+1/2}) \cdot \vec{n}$$

where the last integral is computed with the standard 2 points Gaussian formula. For systems, one easily shows [RCD05, Ric05, DR07] that (4.29) is replaced by

$$\Delta t \leq \min_{K \in \Omega_h} \min_{j \in K} \frac{4}{3} \frac{|K|}{\rho((\vec{a}_{K_t} \cdot \vec{n}_j)^+)}$$

with  $\rho(\cdot)$  the spectral radius of a matrix, and where the positive part of the average Jacobian  $(\vec{a}_{K_t} \cdot \vec{n}_j)^+$  (cf. equation (2.2)) is computed using standard matrix decomposition. We then solve (4.28) discarding the second set of equations. We consider two splittings. One is the first order stMU N scheme obtained by formally generalizing (3.3) (replacing the  $k_i^+$  by the  $\bar{k}_i^+$ , and  $\phi^K$  by  $\Phi^{K_t}$ ). To this stN scheme we can apply the limiting (3.4), (3.5), and (3.6), obtaining the limited stN scheme (LstN).

As a first test we consider a planar  $M_S = 3$  shock in a 50% water-air mixture. The simulations are performed on the mesh on the left on figure 4.8, with periodic boundary conditions in the  $y$  direction [Ric05]. Results after the shock has travelled  $1m$  are shown on figure 4.7, confirming monotone shock capturing and conservative character of the schemes. Note the sharp capturing obtained with the LstN scheme.

The second test involves the interaction of a planar  $M_S = 3$  shock traveling in a mixture containing 80% air with a stationary planar contact discontinuity across which the content of air jumps to 95%. A sketch of the initial state is reported on the right on figure 4.8. A reference solution is computed on a one dimensional grid with a first order FV scheme (simple upwind flux splitting [138]) on 20000 cells.

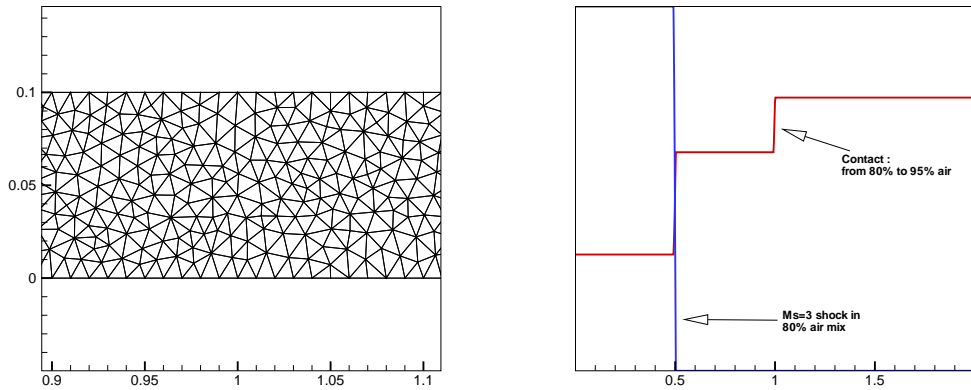


Figure 4.8: Left : mesh for pseudo one-dimensional tests. Right : shock interacting with contact, sketch of the initial state

On figure 4.9 we report the solution obtained with the limited stN scheme on the mesh of figure 4.8 (data extracted along the middle of the domain, line  $y = 0.05$ ). We can see that the contact is set into motion when crossed by the shock and that an intermediate state appears (with roughly 90% air). Across this discontinuity, the pressure remains perfectly constant, even if no particular attention has been given to prevent it to oscillate. This is probably due to the weakness of the jump in volume fraction. The only numerical artifact visible is the start up error, visible in the small bump roughly at  $x = 0.25$ . This is a known phenomenon in shock capturing methods [281, 153, 214, RCD05]. Aside from this effect, no spurious oscillations are observed : the capturing of the discontinuities is clean and sharp.

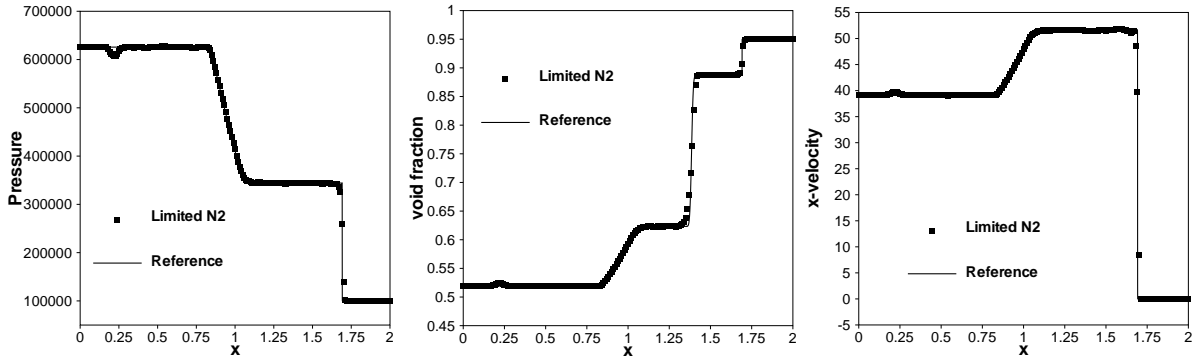


Figure 4.9: Two-phase  $M_S = 3$  shock interacting with a stationary contact. Results of the LstN scheme (denoted as Limited N2) Pressure (left), void fraction (middle), and x-velocity (right). Data extracted along the line  $y = 0.05$

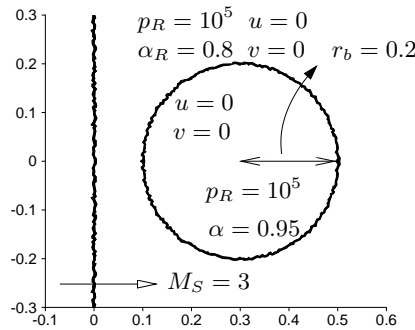


Figure 4.10: Two-phase  $M_S = 3$  shock interacting with a circular contact : initial solution.

Lastly, we compute a two-dimensional version of the same interaction, obtained by replacing the planar stationary contact by a circular one. The test reproduces the interaction of a shock traveling in a 80% air mixture with a bubble containing 95% air. A sketch of the initial solution is reported on the left, and numerical Schlieren visualizations based on the norm of the gradient of the mixture density obtained with the limited stN scheme are reported on figure 4.11. The mesh used has the same topology of the one on figure 4.8. The mesh size is  $h = r_b/40$ ,  $r_b$  being the radius of the bubble.

We can see how the shock sets the contact into motion, and the subsequent roll up of the density discontinuity.



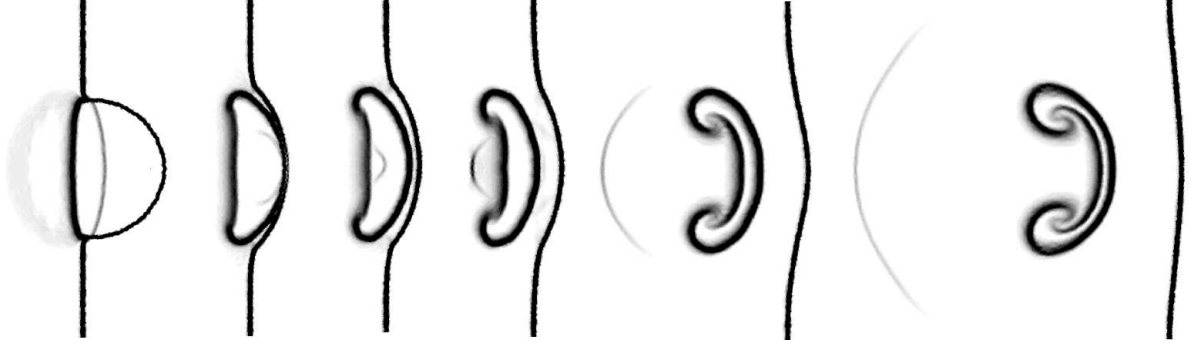


Figure 4.11: Two-phase  $M_S = 3$  shock interacting with a circular stationary contact. Numerical Schlieren (gradient of the mixture density). Result obtained with the LstN scheme.

### 4.2.3 Space-time schemes with discontinuous representation in time

A different approach toward unconditionally positive high order schemes is proposed in [HR11, HRS11]. The idea is to introduce a discontinuous representation of the unknown in time. With reference to figure 4.12, every node of the mesh is represented by its values at time  $t^n-$  and  $t^n+$  (squares and circles in the figure), the two being *a priori* different. One way to present the scheme is to introduce the additional fictitious cells  $K^- = K \times [t^{n-}, t^{n+}]$ , and compute additional space-time residuals on these singular cells. Using standard notation for the jump of  $u_h$ :  $[u_h^n] = u_h(t^{n+}) - u_h(t^{n-})$ , and denoting by  $|K|\overline{\varphi}_j$  the integral over  $K$  of the  $j$ th (spatial) shape function, we have :

$$\Phi^{K^-} = \int_{t^{n-}}^{t^{n+}} \int_K (\partial_t u_h + \nabla \cdot \mathcal{F}_h(u_h)) = \int_K [u_h^n] = \sum_{j \in K} |K| \overline{\varphi}_j (u_j^{n+} - u_j^{n-}) \quad (4.35)$$

Several arguments can be used to justify the fact that no fractions of  $\Phi^{K^-}$  are distributed to the nodes at  $t^{n-}$ . The one more in line with the discussion so far is to say that any space time MU scheme applied to a layer of cells of width  $t^{n+} - t^{n-} \rightarrow 0$  will not distribute to  $t^{n-}$ .

We obtain a new discrete model given by (BCs are not included, and cf. equation (4.28)) :

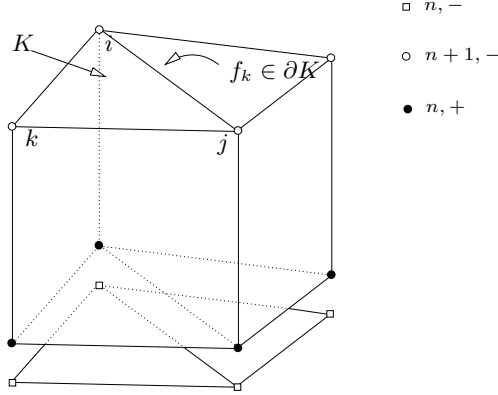
$$\begin{aligned} \sum_{K \in K_i} \overline{\Phi}_i^{K_t} &= 0 \quad \text{for node } i \text{ at } t^{n+1,-} \\ \sum_{K \in K_i} (\widehat{\Phi}_i^{K_t} + \Phi_i^{K^-}) &= 0 \quad \text{for node } i \text{ at } t^{n,+} \end{aligned} \quad (4.36)$$

with of course

$$\sum_{j \in K} (\overline{\Phi}_j^{K_t} + \widehat{\Phi}_j^{K_t}) = \Phi^{K_t} \quad \text{and} \quad \sum_{j \in K} \Phi_j^{K^-} = \Phi^{K^-}$$

where now  $\Phi^{K_t}$  only uses the values  $u_h(t^{n+})$  and  $u_h(t^{n+1,-})$ . Concerning the definition of the splitting of  $\Phi^{K^-}$ , a simple approach allowing to retain second order of accuracy is suggested by the truncation error analysis (details are omitted), in particular we set :

$$\Phi^{K^-} = \int_K \psi_i [u_h^n] \overset{\text{Trapezium rule}}{\approx} \sum_{j \in K} \frac{|K|}{3} \psi_i(\vec{x}_j) [u_j^n] = \frac{|K|}{3} [u_i^n]$$

Figure 4.12: Numerical discretization : space-time element  $K_t$  with time discontinuity

Within the prism, we use any known positive high order nonlinear distribution. As discussed in [HR11], in this case we can write for a scalar problem

$$u_i^{k+1} = u_i^k - \bar{\omega}_i \sum_{j \in K_t} \bar{c}_{ij} (u_i^k - u_j^k) \quad \text{for node } i \text{ at } t^{n+1,-}$$

$$u_i^{k+1} = u_i^k - \hat{\omega}_i \sum_{j \in K_t} \hat{c}_{ij} (u_i^k - u_j^k) - \hat{\omega}_i |C_i| (u_i^k - u_i^-) \quad \text{for node } i \text{ at } t^{n,+}$$

if, for example, the equations are solved by means of the pseudo time procedure (4.26). If  $\bar{c}_{ij}, \hat{c}_{ij} \geq 0$ , and provided that  $\bar{\omega}_i \sum_j \bar{c}_{ij} \leq 1$  and  $\hat{\omega}_i (|C_i| + \sum_j \hat{c}_{ij}) \leq 1$ , then  $\forall k > 0$

$$\min_{K \in K_i} \min_{j \in K_t} u_j^k \leq u_i^{k+1} \leq \max_{K \in K_i} \max_{j \in K_t} u_j^k \quad \text{for node } i \text{ at } t^{n+1,-}$$

$$\min \left( u_i^-, \min_{K \in K_i} \min_{j \in K_t} u_j^k \right) \leq u_i^{k+1} \leq \max \left( u_i^-, \max_{K \in K_i} \max_{j \in K_t} u_j^k \leq u_i^{k+1} \right) \quad \text{for node } i \text{ at } t^{n,+}$$

independently on the time step size. More formally we can prove the following.

**Proposition 4.2.3** (Space time schemes - discrete maximum principle). *Provided that*

1.  $\bar{\Phi}_i^{K_t} = \sum_{j \in K_t} \bar{c}_{ij} (u_i(t^{n+1,-}) - u_j)$ , with  $\bar{c}_{ij} \geq 0 \forall \Delta t > 0$
2.  $\hat{\Phi}_i^{K_t} = \sum_{j \in K_t} \hat{c}_{ij} (u_i(t^{n,+}) - u_j)$ , with  $\hat{c}_{ij} \geq 0 \forall \Delta t > 0$
3.  $\Phi_i^{K^-} = \sum_{j \in K} c_{ij}^- (u_i(t^{n,+}) - u_j(t^{n,-}))$ , with  $c_{ij}^- \geq 0$  and  $c_{ij}^- > 0$  for at least one  $j \in K$  ;

the solution of scheme (4.36) verifies the discrete inequality

$$u_-^n = \min_{j \in \Omega_h} u_j(t^{n,-}) \leq u_i(t^{n,+}), u_i(t^{n+1,-}) \leq \max_{j \in \Omega_h} u_j(t^{n,-}) = U_+^n$$

*Proof.* The proof is obtained by rewriting (4.36) as

$$C U = b^-$$

where, if  $N_{\text{tot}}$  is the total number of nodes in the mesh,  $C$  is a  $2N_{\text{tot}} \times 2N_{\text{tot}}$  matrix,  $U$  contains all the nodal vales at time  $t^{n,+}$  and  $t^{n,+1-}$ , and  $b^-$  is a  $2N_{\text{tot}}$  array containing zeros in the first  $N_{\text{tot}}$  entries, and the right hand sides  $c_{ij}^- u_j(t^{n,-})$  in the remaining  $N_{\text{tot}}$ .

To prove the proposition we note that by hypothesis

- $C$  is an L-matrix ( $C_{ii} \geq 0$ ,  $C_{ij} \leq 0$ ) ;
- $C$  is irreducibly diagonally dominant. In particular, the rows from  $N_{\text{tot}} + 1$  to  $2N_{\text{tot}}$  are easily shown to verify :

$$|C_{ii}| - \sum_j |C_{jj}| = \sum_{K \in K_i} \left( \sum_{j \in K_i} \hat{c}_{ij} + \sum_{j \in K} c_{ij}^- \right) - \sum_{K \in K_i} \sum_{j \in K_i} \hat{c}_{ij} = \sum_{K \in K_i} \sum_{j \in K} c_{ij}^- > 0$$

due to hypothesis 3.

As a consequence  $C$  is an irreducibly diagonally dominant L-matrix, and its inverse is a positive matrix [25] :  $C_{ij}^{-1} \geq 0 \forall i, j$ .

We now recast the right hand side as  $b^- = C^- U^-$  where  $C^-$  is the  $2N_{\text{tot}} \times 2N_{\text{tot}}$  containing zeros everywhere, except in the lower block  $[N_{\text{tot}} + 1, 2N_{\text{tot}}] \times [N_{\text{tot}} + 1, 2N_{\text{tot}}]$  in which  $C_{ij}^- = c_{ij}^-$ . Not also that

$$C\mathbf{1} = C^- \mathbf{1} = r^- \quad (4.37)$$

where  $\mathbf{1}$  is the  $2N_{\text{tot}}$  vector of ones,  $r^-$  contains zeros in the first  $N_{\text{tot}}$  entries, and for  $i \geq N_{\text{tot}} + 1$  :  $r_i^- = \sum_{K \in K_i} \sum_{j \in K} c_{ij}^- > 0$ . Finally (equalities/inequalities meant by component)

$$CU = C^- U^- \stackrel{C_{ij}^- \geq 0}{\geq} C^- \mathbf{1} u_n^- \stackrel{\text{eq. (4.37)}}{=} C\mathbf{1} u_n^-$$

The left inequality is obtained upon multiplication by the positive matrix  $C^-$ . Similarly one obtains the right inequality.  $\square$

The name of the game is to construct schemes that verify the hypotheses 1. and 2. This is readily obtained by using formal extentions of positive schemes for steady computations such as the space time variant of the N scheme (stN) or of the LF scheme (2.39).

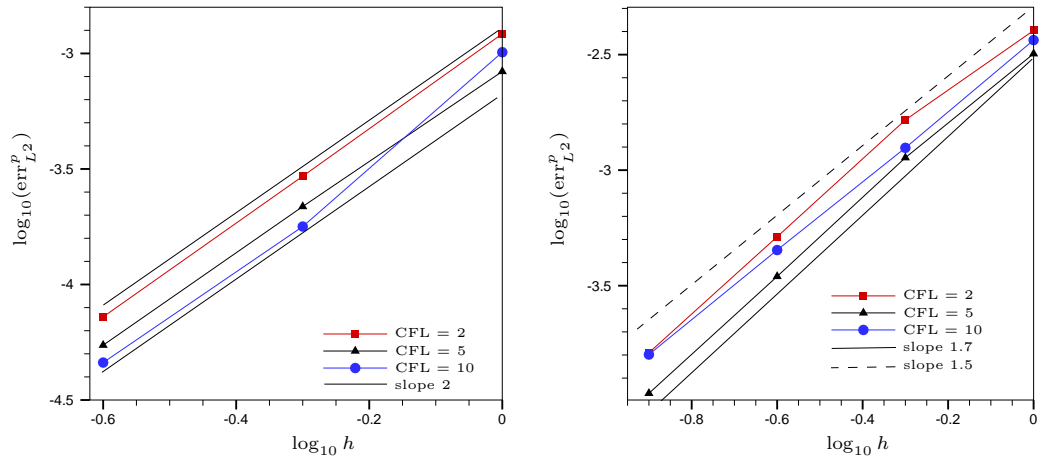


Figure 4.13: Grid convergence for the vortex advection problem :  $L^2$  pressure error for the LDA scheme (left);  $L^2$  pressure error for the LDA-N scheme (right).

We report some examples taken from [HR11]. The results are obtained with space time variants of the multidimensional N, LDA, and blended LDAN schemes of section §2.2.5, recast in conservative form using the approach discussed in paragraph §3. Equations (3.3) and (2.41) are applied, replacing the steady state residual by the  $\Phi^{K,t}$ , and the steady inflow parameters by the space time ones (cf. equation (4.27)). Extension to the system of the Euler equations is achieved with the matrix formulation of [256]. We have set

$$\Delta t = \text{CFL} \min_{K \in \Omega_h} \min_{j \in K} \frac{4}{3} \frac{|K|}{\rho(|\vec{a}_K \cdot \vec{n}_j|)} \quad (4.38)$$

with the average flux Jacobian  $\vec{a}_K$  (cf. equation (2.2)) evaluated at  $t^{n-}$ .

First, the accuracy of the LDA and of the and LDA-N is studied by solving a vortex advection problem. We refer to [96] for a description of the test. Grid convergence plots for different CFL values are reported in Figure 4.13, in which the  $L^2$  norm of the pressure error is plotted. The results show that, for the Euler equations, second or almost second order accuracy is still achieved for all CFL numbers. Clearly, the nonlinear scheme has a much larger error and the slopes can be improved. This shows that better definitions of the blending (or other improved definitions the nonlinear scheme) parameter have to be investigated.

To compare with the results of [RCD01, CRDP01] we have run the supersonic wind tunnel test of [274], on the same mesh shown on figure 4.4. As in [RCD01, CRDP01], we have chosen the time step to compensate for the stiffness introduced by the mesh refinement at the corner. This is achieved by setting  $\text{CFL} = h_{\max}/h_{\min}$ . The results show a nice capturing of the flow features, with perfect monotone shock capturing.

### 4.3 Schemes based on implicit time-stepping

A different approach to construct high order scheme for time dependent problems is explored in [MRAD03, RCD04, RCD05, RA06, RB09b]. The idea is to exploit directly the framework of the accuracy analysis of section §4.1. In particular, let  $\Gamma^{n+1}(u)$  be given by (4.2), we consider schemes that, given  $\{u_h^{n+1-l}\}_{l=1}^{\max(p+1,q)}$  compute the nodal values of  $u_h^{n+1}$  by solving (BCs are omitted) :

$$\sum_{K \in K_i} \Phi_i^K = 0, \forall i \in \Omega_h \quad (4.39)$$

where now  $\forall K \in \Omega_h$  (cf. equation (4.2))

$$\sum_{j \in K} \Phi_j^K = \Phi^K = \int_K \Gamma^{n+1}(u_h) = \int_K \left( \sum_{i=0}^p \alpha_i \frac{\delta u_h^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \nabla \cdot \mathcal{F}_h^{n+1-j} \right) \quad (4.40)$$

where we recall that the weights  $\alpha_i, \theta_j$  are associated to a multi-step time integration scheme.

Formalism (4.39)-(4.40) encompasses also continuous finite element discretizations of the SUPG type combined with high order time integration [48, 50, 49, 29, 30]. In the RD context, the first examples fitting this formalism are due to [174] and [107] who actually used a finite element analogy, and to Doru Caraeni [54, 52, 53] who combined second order backward differencing in time with the LDA distribution of  $\Phi^K$ . The developments that followed tried to improve by providing some tools to construct high order and positivity preserving schemes.

The first developments in this direction are in the PhD of M. Mezine (supervisor R. Abgrall) [180, 9]. Even though presented as a space-time scheme, the discretization proposed

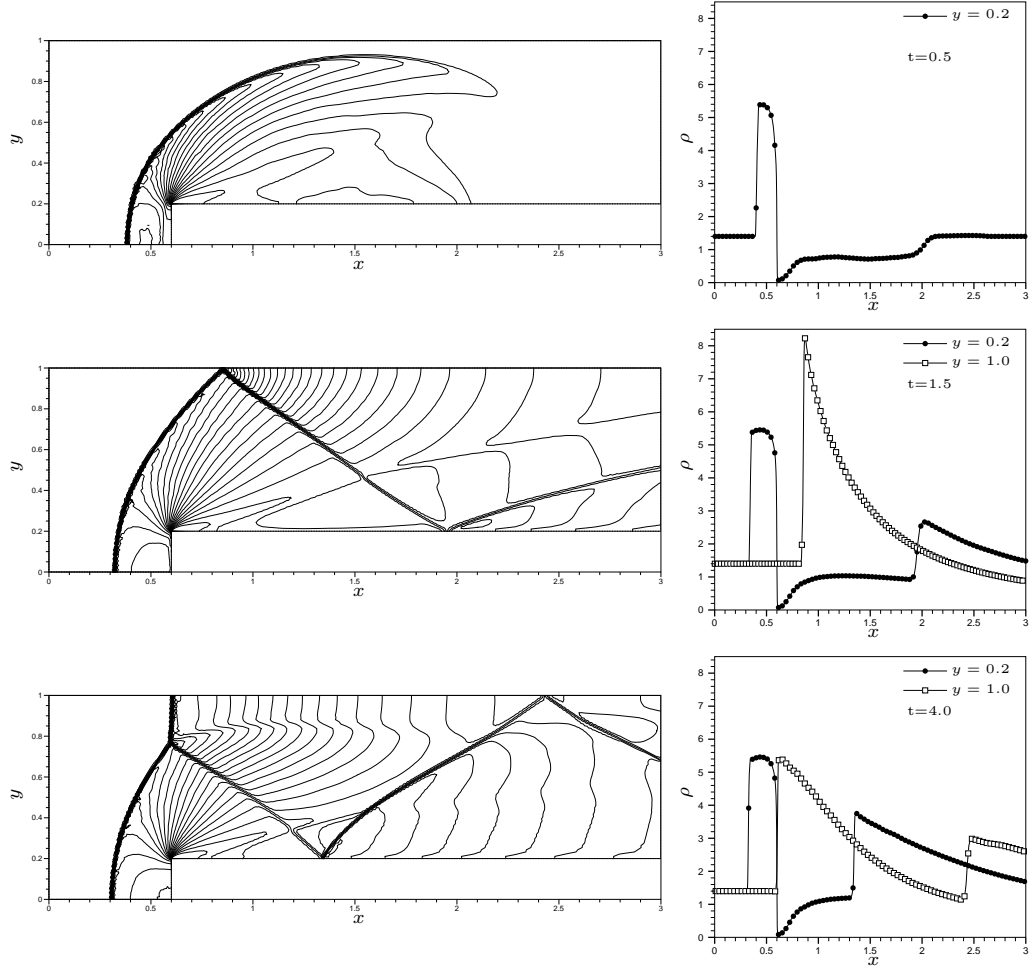


Figure 4.14: Mach 3 facing step test case. Density contours, and density distribution along the lines  $y = 0.2$  (containing the corner singularity), and  $y = 1.0$  (upper wall)

fits more in the context of schemes using a time stepping strategy. In the references, the authors define for  $u_h$  linear in space and time, and for  $\mathcal{F}_h$  linear in time :

$$\Phi^K = \int_{t^n}^{t^{n+1}} \int_K (\partial_t u_h + \nabla \cdot \mathcal{F}_h) = \sum_{j \in K} \frac{|K|}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \int_K \nabla \cdot \mathcal{F}_h^n + \frac{\Delta t}{2} \int_K \nabla \cdot \mathcal{F}_h^{n+1} \quad (4.41)$$

They propose a first order scheme based on the standard N scheme (2.40) defined by

$$\Phi_i^N = \frac{|K|}{3} (u_i^{n+1} - u_i^n) + \frac{\Delta t}{2} (u_i^n - u_{in}^n) + \frac{\Delta t}{2} (u_i^{n+1} - u_{in}^{n+1}) \quad (4.42)$$

Last definition indeed is a splitting of the residual in the sense that  $\sum_j \Phi_j^N = \Phi^K$ , however

when assembling the complete discrete equations on gets

$$|C_i|(u_i^{n+1} - u_i^n) + \frac{\Delta t}{2} \sum_{K \in K_i} (u_i^n - u_{\text{in}}^n) + \frac{\Delta t}{2} \sum_{K \in K_i} (u_i^{n+1} - u_{\text{in}}^{n+1}) = 0$$

which is nothing else than (2.11) for the N scheme, with the trapezium scheme in time replacing Explicit Euler.

So far, the only positive scheme fitting the the continuous in time space time framework of section §4.4.4 was the stMU N scheme using the space time Jacobians  $\widehat{k}$  (cf. equation (4.27)) for upwinding (in space and time simultaneously). No other scheme could ensure the decoupling of the space time slabs. The intuition in the N scheme (4.42) proposed by [9], is that a positive scheme that defines a splitting of (4.41) is obtained simply by considering a positive scheme for the steady problem and integrating it with the trapezium scheme in time. This somehow unifies the  $P^1$  space time approach based on space time multidimensional upwinding, and schemes based on implicit time integration, in particular trapezium or Crank Nicholson.

Unfortunately, this interpretation does not give a means of constructing an implicit unconditionally positive high order scheme. The reason for this is that any high order time integration scheme will preserve the positivity (or the monotonicity [32]) only under a time step limitation [32, 119]. The only unconditionally positivity preserving scheme is the first order implicit Euler scheme. In particular, the following result can be easily proved [DR07, Ric05].

**Proposition 4.3.1** ( $\theta$  scheme and discrete maximum principle). *Let  $\phi_i^P$  denote the splitting of a positive linear first order scheme for steady advection ;*

1.  $\sum_{j \in K} \phi_j^P = \phi^K = \sum_{j \in K} k_j u_j = \int_K \vec{a} \cdot \nabla u_h$
2.  $\phi_i^P = \sum_j c_{ij}(u_i - u_j)$ ,  $c_{ij} \geq 0$

Then, upon integration in time with the  $\theta$ -scheme

$$|C_i|(u_i^{n+1} - u_i^n) + \theta \Delta t \sum_{K \in K_i} \phi_i^P(u_h^{n+1}) + (1 - \theta) \Delta t \sum_{K \in K_i} \phi_i^P(u_h^n) = 0 \quad (4.43)$$

the global discrete maximum principle

$$u_{\min}^n = \min_{j \in \Omega_h} u_j^n \leq u_i^{n+1} \leq \max_{j \in \Omega_h} u_j^n = U_{\max}^n$$

holds under the time step restriction

$$|C_i| \geq (1 - \theta) \Delta t \sum_{K \in K_i} \sum_j c_{ij}$$

In particular, for  $\theta = 1$  (Implicit Euler) the scheme is unconditionally positive

The importance of linear first order schemes is related to the possibility of using them to construct a positive nonlinear high order one. As discussed in section §2.2.6, this can be done either via some blending, or upon application of limiter (2.44). In both cases, a local positivity condition for these constructions to be well defined. This condition boils down to

the requirement that in (4.43) the coefficient multiplying  $u_i^{n+1}$  should be positive, and all the others should be negative, so that the scheme is equivalent to a linear system

$$AU^{n+1} = BU^n$$

with  $A$  an L-matrix (possibly irreducibly diagonally dominant) and  $B$  a positive matrix. However, using the fact that

$$A = \sum_K A^K, \quad B = \sum_K B^K$$

one obtains a local positivity condition by requiring  $A^K$  to be an L-matrix and  $B^K$  to be a positive one. This can be shown to lead to [DR07, Ric05]

**Proposition 4.3.2** (Local positivity,  $\theta$ -scheme). *Under the hypotheses of proposition 4.3.1, a sufficient condition for the  $\theta$ -scheme to verify the discrete maximum principle is*

$$\frac{|K|}{3} \geq (1 - \theta)\Delta t \sum_{j \in K} c_{ij}$$

**Remark 4.3.3** (N and stN schemes vs time step). *Surprisingly, when applying this condition to the N scheme (4.42) proposed in [9], we obtain exactly the past shield condition (4.29) !! However, the two conditions are completely different in nature :*

- *The stN scheme would be unconditionally positive, however the satisfaction of (4.29) is necessary to guarantee the time marching character of the scheme ;*
- *The N scheme is obtained by applying a time marching scheme, and the satisfaction of (4.29) is a sufficient (although not necessary) condition for the satisfaction of a discrete maximum principle*

### 4.3.1 Nonlinear schemes : survey and comparison

This unification has been discussed in [RCD04, RCD05] (see also [Ric05, DR07]), where a thorough comparison between N, stN schemes, and limited high order variants is presented.

In this sub-section we want to report and complete the survey. In particular, we consider here scheme (4.39)-(4.40) with trapezium rule integration in time. For a given linear first order scheme respecting a discrete maximum principle, we consider the nonlinear scheme obtained by the application of limiter (3.4)-(3.5)-(3.6). In view of the comparison on system of equations we summarize hereafter the procedure used in the simulations.

1.  $\forall K$  compute the residual (4.41). The standard 2 points Gauss integration formula is used on each edge of the triangles ;
2. Compute linear first order splittings. The schemes used are :
  - stN scheme coupled with the conservative formulation of paragraph §3.1 (cf. equation (3.3))

$$\Phi_i^{\text{stN}} = \widehat{k}_i^+ (u_i^{n+1} - \widehat{u}_c), \quad \widehat{u}_c = \left( \sum_{j \in K} \widehat{k}_j^+ \right)^{-1} \left( \sum_{j \in K} \widehat{k}_j^+ u_j - \Phi^K \right)$$

- N scheme of [9] coupled with the conservative formulation of paragraph §3.1 (see equation (3.3) for the definition of  $u_c$ )

$$\Phi_i^N = \frac{|K|}{3}(u_i^{n+1} - u_i^n) + \frac{\Delta}{2}k_i^+(u_i^n - u_c^n)\frac{\Delta}{2}k_i^+(u_i^{n+1} - u_c^{n+1})$$

- LF scheme with trapezium rule in time

$$\Phi_i^{\text{LF}} = \frac{|K|}{3}(u_i^{n+1} - u_i^n) + \frac{\Delta t}{3} \oint_K \frac{\mathcal{F}_h^n + \mathcal{F}_h^{n+1}}{2} \cdot \vec{n} + \Delta t \frac{\alpha}{3} \sum_{j \in K} (u_i^{n+1/2} - u_j^{n+1/2})$$

3. Compute nonlinear high order splittings by applying the limiter (3.4)-(3.5)-(3.6). The schemes obtained are referred to as the LstN, LN and LLF schemes respectively ;

- 3.a Following the constructions discussed in [RA06, RB09b] add a streamline dissipation term to the LLF scheme to eliminate spurious modes. The motivation, the analysis *and* the construction behind this step are very similar to those discussed in section §3.2.3. We omit the details. The final modification is

$$\Phi_i^{\text{LLFs}} = \Phi_i^{\text{LLF}} + \delta(u_h) \frac{k_i}{|K|} \tau \Phi^K$$

LLFs standing for Limited Stabilized LF. We refer to [RA06, RB09b] for more details, including the definition of the sensor  $\delta(u_h)$ , and of the scaling matrix  $\tau$  ;

4. Solve the nonlinear algebraic problem (4.39).

In all the computations, the time step is obtained by imposing (4.38) with CFL=1, which is slightly stricter than the past-shield/local positivity condition for the N schemes, while being an approximation of the positivity condition of the LF scheme.

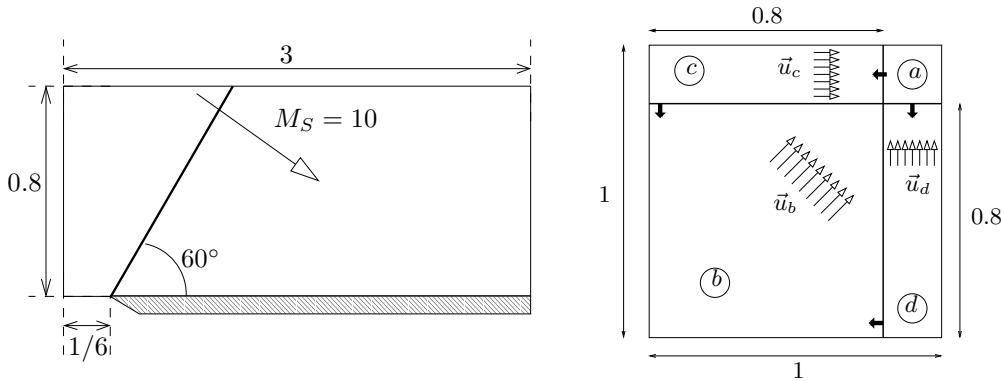


Figure 4.15: Double Mach reflection (left) and shock-shock interaction (right)

We will discuss the comparison on two problems involving the solution of the Euler equations for a perfect gas, on the shock circular contact discontinuity problem of figure 4.11, and give one example of a result obtained with a different time stepping scheme.

The first case considered is the well known double Mach reflection of a planar Mach 10 moving shock on a ramp. A sketch of the problem is reported on the right on figure4.15. As



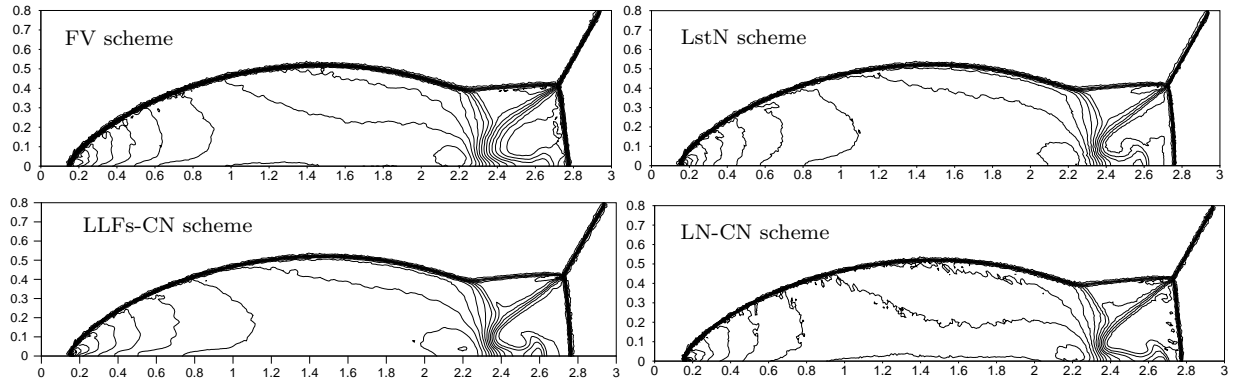


Figure 4.16: Double Mach reflection : density contours. Cell-centered  $\mathcal{FV}$  scheme (top-left), LstN scheme (top-right), LLFs-CN scheme (bottom-left), and LN scheme (bottom-right)

seen in the picture, and as done in [274], the problem is solved on a rotated domain aligned with the ramp. Solutions are computed on an unstructured triangulation with the topology shown on figure 4.8, and with size  $h = 1/100$ .

We compare on figure 4.16 the contours of the density obtained on the same mesh with the three nonlinear RD schemes, and with the cell centered finite volume (FV) scheme with least squares reconstruction, Roe flux, the limiter of Barth and Jespersen [23], and a second order TVD explicit Runge Kutta time integrator [119], with time step given by (4.38) with CFL= 1/4. First of all, the results show the shock capturing capabilities of the RD schemes : no oscillations are visible, and both the contact emanating from the triple point, and the jet of material on the ramp are well resolved. The main difference in the results is observed precisely in the resolution of these multidimensional features. The worst result is obtained with the FV scheme which yields a very thick contact and a poorly resolved jet. The best result is the one obtained with the LN scheme which shows a very crisp resolution of the contact and of the jet which is stronger than in the other results (closer to the normal reflection). The second best is the LLFs scheme, which is also by far the simplest among the RD schemes. The LstN scheme is still better than the FV scheme which, however, has on its side the fact of being genuinely explicit, thus the fastest of the four to run.

The next test is a shock-shock interaction proposed in [163] and studied by several others to asses the capability of a scheme to capture truly multidimensional wave interactions (see *e.g.* [170]). A sketch of the initial state is given on the right on figure 4.15 (see [163, 170, RCD05] for a quantitative description) : two normal shocks and two oblique ones converge in one point ; due to the interaction four irregular reflections appear and a jet of material is pushed in the low pressure region between the two oblique shocks From two of the triple points strong contacts emanate and interact with one of the shock legs of the other reflection.

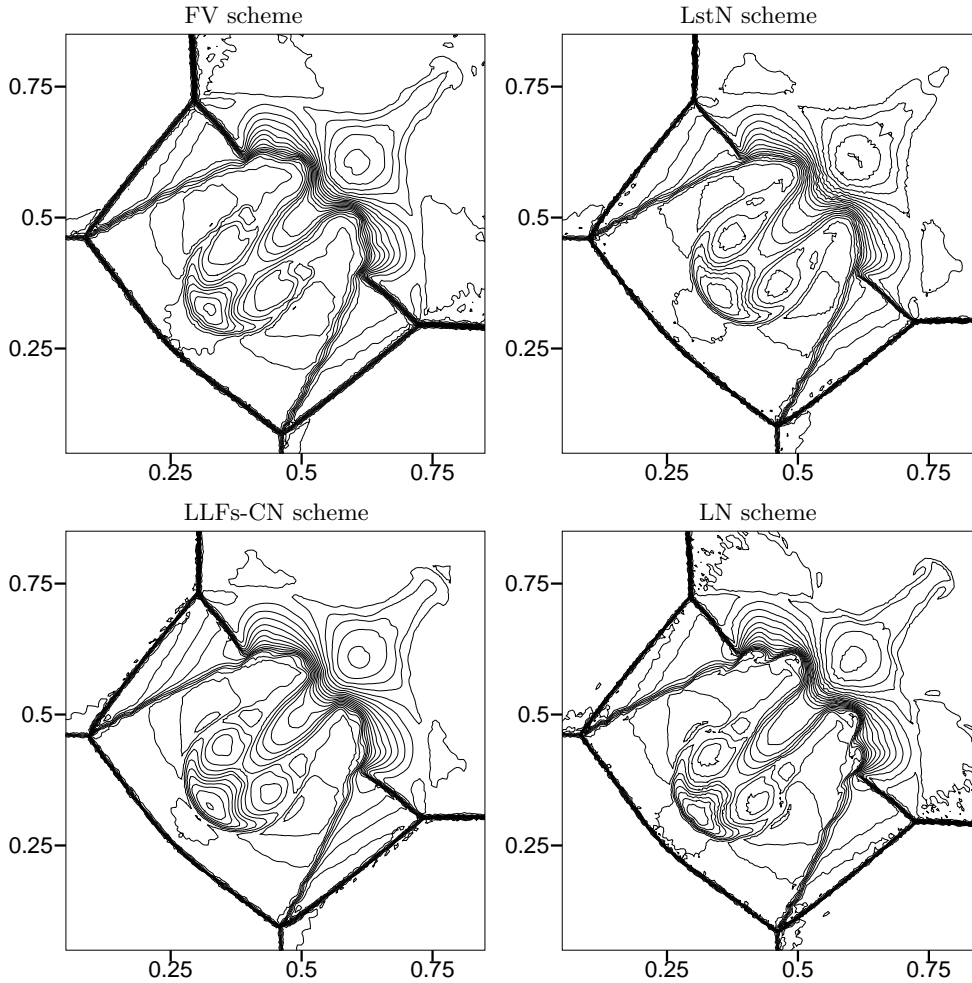


Figure 4.17: Shock-shock interaction. Contours of the density obtained with the LST-N scheme (top),  $\mathcal{FV}$  scheme (bottom-left) and LN scheme (bottom-right)

This very complex interaction is visible in the density contours reported on figure 4.17, where the results of the three nonlinear RD schemes and of the FV one. As in the double Mach reflection case, the solutions are free of oscillations, and all the features of the interaction are nicely reproduced. Also in this case, the LN scheme yields very crisp contact discontinuities, and even a *glimpse* of a Kelvin-Helmholtz instability. Note that, even if this type of instability is physical (and inviscid), the trigger is of course unphysical and related to numerical perturbations (even spurious modes). The numerical evolution computed with the Euler equations is also unphysical, a correct approximation requiring the solution of the full Navier-Stokes equations. However, its visibility in the LN scheme results is a sign of low numerical dissipation. Among the other schemes, the worst is definitely the LstN scheme. The LLFs scheme yields thinner contacts and a better resolved jet (even though though slower) than the FV scheme. Again, the net advantage of the FV scheme is its explicit character.

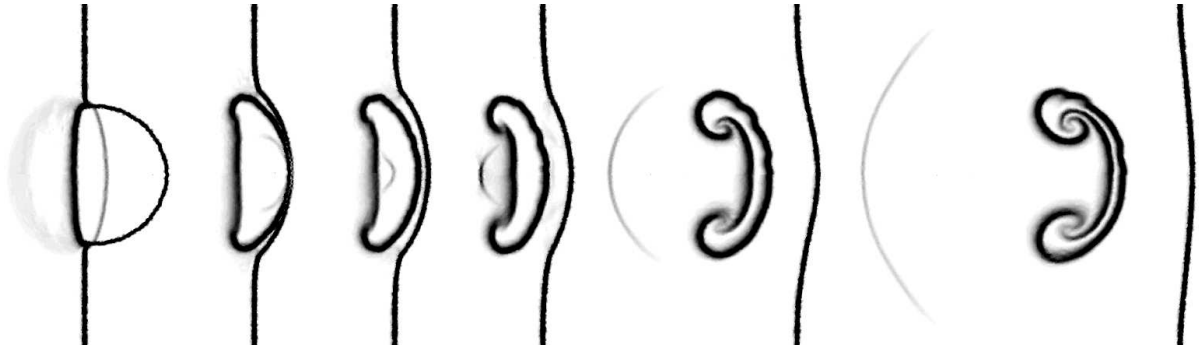


Figure 4.18: Two-phase  $M_S = 3$  shock interacting with a circular stationary contact. Numerical Schlieren (gradient of the mixture density). Comparison between the LN scheme (top halves) and the LstN scheme (bottom halves).

To compare the conservative formulations of the N and stN schemes, we report the result obtained on the interaction of a planar  $M_S = 3$  shock traveling in a air-water mixture with 80% air, with a circular contact enclosing a region of pure air (cf. section §4.2.2). We report on figure 4.18 a comparison of the results obtained with the LstN and LN schemes in terms of numerical Schlieren images obtained from the gradient of the mixture density. In all the figures, the lower halves contain the result of the LstN scheme (same as on figure 4.11), while the top halves being the images obtained from the density computed by the LN scheme.

The evolution is very similar, and no differences are observed in the initial interaction. However, as the shock traverses the circular *bubble* we clearly see that the LN scheme gives a crisper resolution of the density interface. The evolution of this interface is richer of details in the LN solution, the contact showing again a glimpse of an inviscid instability (wavy shape of the mushroom head). We refer to [RCD05, Ric05, DR07] for more results and comparisons.

Lastly, we present an example obtained with a different time stepping approach. In particular, we consider the LLFs distribution in conjunction with the standard second order backward difference (Bdf2) scheme [119, 32] obtained by setting in (4.40)

$$q = 0, \theta_0 = 1 \quad i = 1, \alpha_0 = \frac{3}{2}, \alpha_1 = -\frac{1}{2}$$

We solve the shallow water equations (1.4) on an unstructured triangulation of the square  $[0, 100]^2$  with mesh size  $h = 2$  and initial solution  $\vec{v} = 0$  and

$$d = \begin{cases} 10 & \text{if } r \leq 60 \\ 0.5 & \text{otherwise} \end{cases}$$

with  $r$  the distance from the origin. We show the results in terms contours and one dimensional profiles of free surface and of the Froude number

$$\text{Fr} = \frac{\|\vec{v}\|}{\sqrt{gd}}$$

We recall that the Fr number plays in shallow water flows the same role of the Mach number is gas dynamics. Since the Bdf2 scheme has no positivity preservation properties [32], we *blindly* set CFL= 2 in (4.38). The results obtained are reported on figure (4.19).

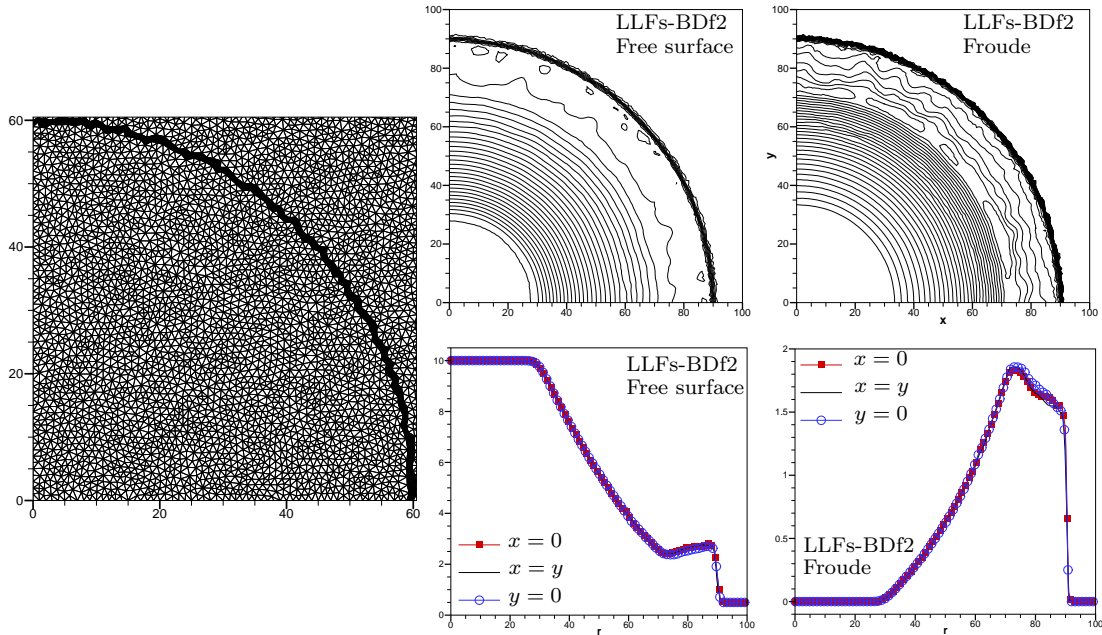


Figure 4.19: Trans-critical circular dam break : Limited LF-BDf2 scheme with stabilization. Free surface and Froude number contours (left), and line plots (right)

As in all previous cases, we observe a nice monotone capturing of the shock (bore), while the contours in the smooth expansion are nicely circular. Note that for the initial condition chosen, the flow becomes super-critical ( $Fr > 1$ ), as it can be seen by the 1d line plots. The contour plots show that in correspondence of the critical point  $Fr=1$  the contours stay smooth, a *sonic* glitch being barely visible.

The irregularities in the Froude contours right before the shock are largely related to the irregular structure of the initial condition in which the circular discontinuity is interpolated on the unstructured triangulation.

These results give some hope for the construction of simple, yet effective schemes for computing discontinuous solutions. Simplicity being related to the use of the LF scheme as the underlying first order discretization, and the effectiveness being related to the possibility of increasing the CFL in (4.38). Unfortunately, this is only true if strict conservation of the positivity is not sought, as we will discuss in Part II of the manuscript.

## 4.4 Genuinely explicit schemes

This last section tries to answer to the question of section §4.1.3 : are we able to produce a genuinely explicit residual scheme ? The answer is of course affirmative. This has been done in [RA10] for second order of accuracy in space, and it is currently being adapted to higher orders of accuracy. This development is interesting in the framework of the residual schemes object of this manuscript, however it has applications in other residual based discretizations, such as the RBC schemes of Lerat and Corre [76], SUPG and other stabilized Galerkin schemes with *residual* stabilization, high order residual based FV schemes [61, 62].

The main motivation is the following :

- As long as we blindly apply the residual approach, we will have to invert a (possibly nonlinear) mass matrix :
- Schemes based on multistep implicit time integration are somewhat simpler but they cannot guarantee unconditional positivity preservation. A CFL= $\mathcal{O}(1)$  time step limitation has to be respected, thus making the schemes inefficient ;
- Space-time schemes are much more promising from this point of view. Unconditional positivity *and* higher orders can be obtained. Moreover, a technique allowing to construct unconditionally positive two-layer schemes starting from a positive linear scheme for the steady problem is contained in [180, 9]. The price to pay is the complication of the implementation, which will pay back only when going to real stiff applications.

Ideally, we would like to have a framework for the construction of truly explicit schemes, at least to replace the expensive Crank Nicholson/Trapezium rule schemes in computations requiring the exact preservation of positivity (cf. part III of the manuscript).

#### 4.4.1 Digression : on RD and mass matrices

One of the first paragraphs of [RA10] is entitled : “Second order RD : the proliferation of mass matrices”. The reason for such a title is that in the known literature on RD schemes, there are at least three different *consistent* formulations, each with a completely different definition of the RD mass matrix *allowing to recover second order of accuracy*.

Let us for the moment consider the linear constant advection problem

$$\partial_t u + \vec{a} \cdot \nabla u = 0 \quad (4.44)$$

and focus on discrete counterparts of (4.44) that, can be written as

$$\sum_{K \in K_i} \left\{ \sum_{j \in T} m_{ij} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right\} = 0 \quad \forall i \in \mathcal{T}_h \quad (4.45)$$

We then introduce the *nodal residuals*

$$\Phi_i^K(u_h) = \sum_{j \in K} m_{ij} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \quad (4.46)$$

and assume the satisfaction of the consistency relation

$$\sum_{j \in K} \Phi_j^K(u_h) = \Phi^K(u_h) = \int_K (\partial_t u_h + \vec{a} \cdot \nabla u_h) \quad (4.47)$$

with

$$\phi^K(u_h) = \int_K \vec{a} \cdot \nabla u_h \quad (4.48)$$

This prototype is meant to be a consistent generalization to the time dependent case of residual distribution. To simplify the discussion, we keep a distinction between the fluctuation (4.48) and the residual (4.47), the latter representing the integral of the whole equation.

As already anticipated there exist different definitions of  $m_{ij}$ , for a given  $\beta_i^K$ , allowing to recover second order of accuracy. The problem here is that we are trying to reverse engineer a mass matrix, which is normally arising from a variational statement, which we do not have.

What are then the conditions that the  $m_{ij}$  coefficients should satisfy ? To simplify the notation, let us set

$$r_h = \partial_t u_h + \vec{a} \cdot \nabla u_h \quad (4.49)$$

The first condition is implicit in the conservation requirement (4.47). The second is that, if the time derivative is constant in space, then, by consistency with the spatial discretization, everything should be distributed in the same direction. The second condition tries to mimic the behavior of

$$\int_K \omega_i^K r_h$$

for  $r_h$  constant in space. The two conditions imply that we should have

$$\sum_{i \in K} m_{ij} = \frac{|K|}{3}, \quad \sum_{j \in K} m_{ij} = |K| \beta_i^K \quad (4.50)$$

*And this is all !* We cannot say more. It is this lack of constraints that allows so many different formulations such as :

**Residual Approach or F1** The method originally proposed by Caraeni [54] :

$$0 = \sum_{K|i \in K}_i \beta_i^K \Phi^H(u_h) = \sum_{K \in K_i} \left( \sum_{j \in K} m_{ij}^{F1} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right), \quad m_{ij}^{F1} = \frac{|K|}{3} \beta_i^K \quad (4.51)$$

with  $\delta_{ij}$  Kroenecker's delta, and F1 standing for Formulation 1.

**Petrov Galerkin approach or F2** This is the method originally proposed by [174, 107] :

$$\begin{aligned} 0 &= \int_{\Omega} \varphi_i r(u_h) + \sum_{K \in K_i} \int_K \delta_{\varphi_i} r(u_h) \\ &= \sum_{K \in K_i} \left( \sum_{j \in K} m_{ij}^{F2} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right), \quad m_{ij}^{F2} = \frac{|T|}{36} (3 \delta_{ij} + 12 \beta_i - 1) \end{aligned} \quad (4.52)$$

with  $\delta_{ij}$  Kroenecker's delta, and F2 standing for Formulation 2, and where for constant  $\delta_{\varphi_i}$ , the second of (4.50) gives immediately  $\delta_{\varphi_i}|_T = \beta_i^K - 1/3$  (in 2d).

**Weighted area Approach or F3** Proposed in [92]. Based on the idea that  $\forall j \in K$  there is a sub-cell  $K^j | j \in K^j$  and  $|K^j| = \beta_j^K |K|^1$  (cf. figure 4.20). This leads finally to

$$\begin{aligned} 0 &= \sum_{K \in K_i} \int_{K^i \subset K} r(u_h) = \sum_{K \in K_i} \left( \sum_{j \in K} m_{ij}^{F3} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right) \\ & \quad m_{ij}^{F3} = \frac{|T|}{3} \beta_i (\delta_{ij} + 1 - \beta_j) \end{aligned} \quad (4.53)$$

with  $\delta_{ij}$  Kroenecker's delta, and F3 standing for Formulation 3.

---

<sup>1</sup>which implicitly assumes  $\beta_i^K \geq 0$

**Weighted area Approach II or F4** Same as in F3, but assuming that  $j \notin K^j$ . Based on the observation that if  $\beta_i^K \geq 0 \forall i$ , we can find a unique point  $M \in K$ , such that  $\varphi_i(M) = \beta_i^K$ . The  $\beta_i$  coefficients represent the area coordinates of  $M$  (right on figure 4.20). With the notation of figure 4.20, we find :

$$0 = \sum_{K \in K_i} \int_{K^i \subset K} r(u_h) = \sum_{K \in K_i} \left( \sum_{j \in K} m_{ij}^{F4} \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right) \quad (4.54)$$

$$m_{ij}^{F4} = \frac{|T|}{3} \beta_i (1 - \delta_{ij} + \beta_j)$$

with  $\delta_{ij}$  Kroenecker's delta, and F4 standing for Formulation 4.

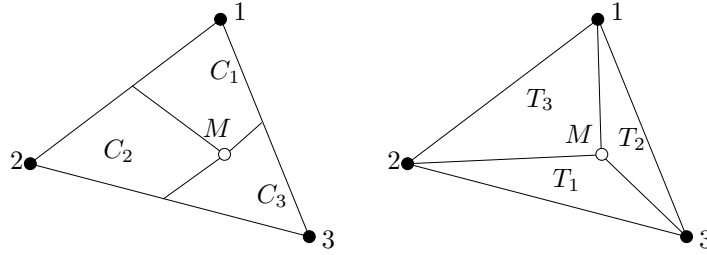


Figure 4.20: Left. Formulation 3 : dual areas  $K^j$ ,  $|K^j| = \beta_j^K |K|$ . Right. Formulation 4 : area coordinates of the distribution point  $M$  ;  $|K|_j = \beta_j^K |K|$

What all these formulations try to do, is mimic

$$\Phi_i^K(u_h) = \int_K \omega_i^K r(u_h) \quad (4.55)$$

with (4.50) becoming

$$\sum_{j \in T} \omega_j = 1, \quad \frac{1}{|T|} \int_T \omega_i dx dy = \beta_i^T \quad (4.56)$$

These two constraints are not enough, and indeed the number of functions that verify these constraints is infinite. Denoting by  $\chi_K$  the characteristic function

$$\chi_K(x, y) = \begin{cases} 1 & \text{if } (x, y) \in K \\ 0 & \text{if } (x, y) \notin K \end{cases}$$

we can easily recover the 4 formulations above by setting

$$\omega_i^{F1} = \sum_{K \in K_i} \beta_i^K \chi_K$$

$$\omega_i^{F2} = \varphi_i + \sum_{K \in K_i} \delta_{\varphi_i} \chi_K$$

$$\omega_i^{F3/F4} = \sum_{K \in K_i} \chi_K$$

Moreover, for any given test function  $\omega_i^K$  verifying all the consistency, conservation, and accuracy constraints, we can easily come up with a modified function, say  $\tilde{\omega}_i$  with all the desirable properties. For example, if we can find three bounded functions, say  $f_1, f_2$ , and  $f_3$  such that

$$\sum_{j=1}^3 f_j = C_f$$

with  $C_f$  a constant, we can modify  $\omega_i^K$  as

$$\bar{\omega}_i = \omega_i^K + k_0(f_i - \bar{f}_i), \quad \bar{f}_i = \frac{1}{|K|} \int_K f_i$$

with  $k_0$  an arbitrary parameter ! In fact, this term neither affects conservation, due to

$$\sum_{j=1}^3 (f_j - \bar{f}_j) = 0, \quad \int_K (f_j - \bar{f}_j) dx dy = 0$$

nor the accuracy, as long as the each  $f_i$  is bounded, nor the consistency with the original distribution, since in the  $P^1$  case

$$\int_T (f_j - \bar{f}_j) \vec{a} \cdot \nabla u_h dx dy = 0$$

so that the extra term only affects the form of the mass matrix.

The last observation leads to some interesting consequences if we take  $f_i = \varphi_i$ . This choice leads to the following modification of the mass matrix :

$$\bar{m}_{ij} = m_{ij} + k_0 \int_K (\varphi_i - \bar{\varphi}_i) \varphi_j$$

that leads to the semi-discrete scheme

$$\sum_{K \in \mathcal{K}_i} \left( \sum_{j \in K} (m_{ij} + k_0 \delta m_{ij}) \frac{du_j}{dt} + \beta_i^K \phi^K(u_h) \right) = 0, \quad \delta m_{ij} = \frac{|K|}{36} (3\delta_{ij} - 1) \quad (4.57)$$

with  $\delta_{ij}$  Kroenecker's delta. As remarked in [RB09b], the matrix  $\delta m_{ij}$  is symmetric, and defines a dissipation operator :

$$v^T [\delta m_{ij}] v \geq 0, \quad \forall v \in \mathbb{R}^3$$

The interesting observation is that if we take  $k_0 = 3$  and apply the modification to the Galerkin scheme we obtain :

$$\bar{m}_{ij} = \overbrace{\frac{|K|}{12} (\delta_{ij} + 1)}^{\text{Gal}} + \overbrace{\frac{|K|}{12} (3\delta_{ij} - 1)}^{3\delta m_{ij}} = \frac{|K|}{3} \delta_{ij}$$

Which is just another way to show that mass lumping for the Galerkin scheme does not reduce the accuracy in the  $P^1$  case but it does introduce a degree of dissipation.



Finally, by comparing (4.51) and (4.52) with (4.57), we realize that

$$m_{ij}^{F2} = m_{ij}^{F1} + \delta m_{ij} \quad (4.58)$$

Similarly, F3 and F1 mass matrices are linked by a very similar relation :

$$m_{ij}^{F3} = m_{ij}^{F1} + \widetilde{\delta m}_{ij}, \quad \widetilde{\delta m}_{ij} = \frac{|K|}{3} (\beta_i^K \delta_{ij} - \beta_i^K \beta_j^K) \quad (4.59)$$

where, *provided that*  $\beta_i^K \geq 0 \forall i$ , then the symmetric matrix  $\widetilde{\delta m}_{ij}$  also defines a dissipation operator. In particular,  $\forall v \in \mathbb{R}^3$  we have

$$v^T [\widetilde{\delta m}_{ij}] v = \frac{|K|}{3} \beta_1^K \beta_2^K (v_1 - v_2)^2 + \frac{|K|}{3} \beta_1^K \beta_3^K (v_1 - v_3)^2 + \frac{|K|}{3} \beta_3^K \beta_2^K (v_3 - v_2)^2 \geq 0$$

A similar relation holds for the last formulation, only this time we have

$$m_{ij}^{F1} = m_{ij}^{F4} + \widetilde{\delta m}_{ij} \quad (4.60)$$

The net result is that *all the formulations are equivalent up to a dissipation term* :

$$\begin{aligned} m_{ij}^{F1} &= m_{ij}^{F4} + \widetilde{\delta m}_{ij} \\ m_{ij}^{F2} &= m_{ij}^{F1} + \delta m_{ij} = m_{ij}^{F4} + \widetilde{\delta m}_{ij} + \delta m_{ij} \\ m_{ij}^{F3} &= m_{ij}^{F1} + \widetilde{\delta m}_{ij} = m_{ij}^{F4} + 2\widetilde{\delta m}_{ij} \end{aligned} \quad (4.61)$$

**Remark 4.4.1.** *In one space dimension, if the spatial discretization is given by the classical 1d upwind scheme, the formulations F1, F3, and F4 are identical. The formulation F2 reduces to the 1D SUPG scheme obtained by setting for the SUPG parameter [151, 232, 131]*

$$\tau = \frac{\Delta x}{2|a|}$$

This analysis is based on the used of simple algebraic arguments related only to the consistency of the discretization. More answers could come *e.g* from a Fourier analysis, which is under way. Of course if we did have a variational statement, additional constraints would be related to the actual stability of the discretization, as in [48, 29, 30, 49].

The objective of the following sections is to show how some of the ideas presented have been used in [RA10] to obtain genuinely explicit schemes.

#### 4.4.2 Step 1 : stabilized Galerkin and explicit RK integration

We consider here discretizations that can be written as

$$\sum_{K \in \Omega_h} \omega_i^K (\partial_t u_h + \nabla \cdot \mathcal{F}_h) = 0 \quad (4.62)$$

for some test function  $\omega_i$  respecting (4.56). We unduly assume that  $\omega_i^K$  can be written as

$$\omega_i^K = \varphi_i + \gamma_i^K \quad (4.63)$$

where the *stabilization operator*  $\gamma_i^K$  can be highly nonlinear.

Following [RA10], we consider explicit Runge-Kutta (RK) schemes, that for

$$u' + f(u) = 0$$

can be written as a sequence of  $m$  steps of the type

$$u^{k+1} - u^n + \Delta t \left( \overbrace{\sum_{l=1}^k a_{kl} f(u^l)}^{f^k} \right) = u^{k+1} - u^n + f^k = 0 \quad (4.64)$$

with  $f^k$  the  $k$ th iteration RK evolution operator. Similarly, for our conservation law, we introduce the semi-discrete  $k$ th iteration RK residual

$$r_h^{\text{RK}(k)} = \frac{u^{k+1} - u^n}{\Delta t} + \nabla \cdot \left( \overbrace{\sum_{l=1}^k a_{kl} (\vec{a} u_h)^l}^{\mathcal{F}_h^k} \right) = \frac{u^{k+1} - u^n}{\Delta t} + \nabla \cdot \mathcal{F}_h^k \quad (4.65)$$

The application of (4.62) to the semi-discrete  $k$ th RK step gives, taking into account (4.63)

$$\sum_{K \in K_i} \int_K \varphi_i r_h^{\text{RK}(k)} + \sum_{K \in K_i} \int_K \gamma_i^K r_h^{\text{RK}(k)} = 0 \quad (4.66)$$

Due to the assumed nonlinear character of  $\gamma_i^K$ , the obtention of  $u_h^{k+1}$  from (4.66) requires the solution of a nonlinear algebraic problem, or at least the inversion of a mass matrix depending nonlinearly on  $\vec{a}$ .

### 4.4.3 Step 2 : inaccurate residuals and stabilization

The idea put forward in [RA10] is that the semi-discrete residual used in the stabilization integrals does not need to be the same as the one used in the Galerkin integrals. Evidence of this fact comes from the predictor multi-corrector techniques in use since many years in SUPG and Least squares schemes [151, 232, 131, 132, 133].

The objective is then to define a modified semi-discrete residual  $\bar{r}_h^{\text{RK}(k)}$  that leaves intact (at least) the accuracy (rate of convergence) of the scheme (and possibly its stability). The advantage in doing this should be that we get rid of the necessity of inverting the mass matrix, so in [RA10] it is proposed that this modified residual should have the form

$$\bar{r}_h^{\text{RK}(k)} = \frac{\bar{\Delta} u^{k+1}}{\Delta t} + \nabla \cdot \mathcal{F}_h^k, \quad \Delta u^{k+1} = \sum_{j=0}^k \alpha_j u^{k-j} \quad (4.67)$$

so that the stabilization component of the scheme becomes genuinely explicit. Note that in the last expression  $u^0 = u^n$ , while the other values are the intermediate RK solutions.

#### The accuracy analysis

What are the conditions on  $\bar{r}_h^{\text{RK}(k)}$  that allow to retain the initial accuracy ? One can prove the following.

**Proposition 4.4.2** (Accuracy and time-stepping). *Consider a  $p$ -th order  $m$ -step RK scheme verifying the truncation error estimate*

$$r^{RK(m)}(w) = \frac{w^{n+1} - w^n}{\Delta t} + \nabla \cdot \mathcal{F}^m(w) = C_{RK} \Delta t^p \quad (4.68)$$

whenever  $w$  is a smooth enough classical solution  $\partial_t w + \nabla \cdot \mathcal{F}(w) = 0$ . Similarly, assume that the modified semi-discrete operator associated to the modified residual  $\bar{r}$  verifies

$$\bar{r}^{RK(m)}(w) = \frac{\bar{\Delta} w^{m+1}}{\Delta t} + \nabla \cdot \mathcal{F}^m(w) = \bar{C}_{RK} \Delta t^l \quad (4.69)$$

For  $w_h$ , a  $p$ -th order accurate continuous polynomial approximation  $w_h = \sum_{i \in \Omega_h} \varphi_i w_i$ , and a smooth function  $\psi \in C_0^1(\Omega)$ , define the truncation error at time  $t^{n+1}$

$$\begin{aligned} \epsilon_{n+1}(w, \psi) = & \left| \int_{\Omega} \psi_h \left( \frac{w_h^{n+1} - w_h^n}{\Delta t} + \nabla \cdot \mathcal{F}_h^m(w_h) \right) + \right. \\ & \left. \sum_{i \in \mathcal{T}_h} \psi_i \sum_{K \in K_i} \int_K \gamma_i^K \left( \frac{\bar{\Delta} w_h^m}{\Delta t} + \nabla \cdot \mathcal{F}_h^m(w_h) \right) \right| \end{aligned} \quad (4.70)$$

The error (4.70) verifies an estimate of the type

$$\epsilon_{n+1}(w, \psi) \leq C h^p$$

provided that

1. the mesh and time step satisfy the regularity requirements

$$C_0 \leq \sup_{K \in \Omega_h} \frac{h^2}{|K|} \leq C_1, \quad C'_0 \leq \frac{\Delta t}{h} \leq C'_1$$

2. the bubble  $\gamma_i^K$  is uniformly bounded

3. the approximate semi-discrete residual verifies hypothesis (4.69) with

$$l \geq p - 1$$

*Proof.* The first part of the proof boils down to showing that

$$\left| \int_{\Omega_h} \psi_h \left( \frac{w_h^{n+1} - w_h^n}{\Delta t} + \nabla \cdot \mathcal{F}_h^m(w_h) \right) \right| \leq C_a h^p \quad (4.71)$$

under the hypotheses of the proposition. This part is identical to what is done reported in the analysis reported beginning of this chapter (cf. section §4.1.2, estimates for terms I and II), and is omitted.

We consider now the term

$$\begin{aligned} I &= \sum_{K \in \Omega_h} \sum_{j \in K} \int_K \psi_j \gamma_j^K \left( \frac{\bar{\Delta} w_h^m}{\Delta t} + \nabla \cdot \mathcal{F}_h^m(w_h) \right) \\ &= \sum_{K \in \Omega_h} \frac{1}{C_K} \sum_{i, j \in K} \int_K (\psi_j - \psi_i) \gamma_j^K \left( \frac{\bar{\Delta} w_h^m}{\Delta t} + \nabla \cdot \mathcal{F}_h^m(w_h) \right) \end{aligned}$$

having used the fact that, since  $\sum_j \omega_j^K = 1 = \sum_j \varphi_j$ , then  $\sum_j \gamma_j^K = 0$ , and  $C_K$  denoting the number of degrees of freedom of  $K$ .

As done in section §4.1.2, for a  $p$ th order accurate approximation  $w_h$ , and  $\mathcal{F}_h(w_h)$  we now consider the estimates [63, 105]

$$|\overline{\Delta}w_h^m - \overline{\Delta}w^m| = \mathcal{O}(h^p), \quad |\nabla \cdot \mathcal{F}_h^m(w_h) - \nabla \cdot \mathcal{F}^m(w)| = \mathcal{O}(h^{p-1})$$

and, using (4.69) we estimate term I as

$$|\text{I}| \leq C \frac{|\Omega|}{h^2} \frac{h^2}{C_K} \|\nabla \psi\|_{L^\infty} h \sup_{K \in \Omega_h} \sup_{j \in K} (C_\alpha h^p \Delta t^{-1} + C_\beta h^{p-1} + \overline{C}_{\text{RK}} \Delta t^l)$$

The regularity assumptions on the mesh and time step size lead to

$$|\text{I}| \leq C' h^p + \overline{C}_{\text{RK}} h^{l+1}$$

This estimate, together with (4.71) yields the desired result.  $\square$

### Examples of inaccurate residuals

The construction of the modified residuals has been done so far for the standard TVD RK2 and RK3 schemes [119]. The form chosen for  $\overline{r}^{\text{RK}(k)}$  makes it easy to find the coefficients  $\alpha_j$  in (4.67) in order to satisfy (4.69). This has led to the following definitions (for  $u' + f(u) = 0$ ).

#### RK2 scheme :

Step 1 :

$$r^0 = \frac{u^1 - u^n}{\Delta t} + f(u^n)$$

$$\overline{r}^0 = f(u^n)$$

Step 2 :

$$r^1 = \frac{u^{n+1} - u^n}{\Delta t} + \frac{f(u^n) + f(u^1)}{2}$$

$$\overline{r}^1 = \frac{u^1 - u^n}{\Delta t} + \frac{f(u^n) + f(u^1)}{2}$$

#### RK3 scheme :

Step 1 :

$$r^0 = \frac{u^1 - u^n}{\Delta t} + f(u^n)$$

$$\overline{r}^0 = f(u^n)$$

Step 2 :

$$r^1 = \frac{u^{n+1} - u^n}{\Delta t} + \frac{f(u^n) + f(u^1)}{4}$$

$$\overline{r}^1 = \frac{u^1 - u^n}{2\Delta t} + \frac{f(u^n) + f(u^1)}{4}$$

Step 3 :

$$r^2 = \frac{u^{n+1} - u^n}{\Delta t} + \frac{1}{6} (f(u^n) + 4f(u^2) + f(u^1))$$

$$\overline{r}^2 = 2 \frac{u^2 - u^n}{\Delta t} + \frac{1}{6} (f(u^n) + 4f(u^2) + f(u^1))$$

#### 4.4.4 Step 3 : mass lumping

So far, after few manipulations, the scheme obtained reads

$$\int_{\Omega_h} \varphi_i \frac{u_h^{k+1} - u_h^n}{\Delta t} - \int_{\Omega_h} \varphi_i \frac{\bar{\Delta} u_h^{k+1}}{\Delta t} = - \sum_{K \in K_i} \int_K \omega_i^K \bar{r}_h^{\text{RK}(k)} \quad (4.72)$$

This nicely shows that the scheme expresses a balance between two errors : one being the different approximations of the time derivative, the other the weighted approximation of the equation, based on the modified residuals.

The last step, consists in lumping the Galerkin integrals on the left hand side. There are several publications on high order mass lumping for Galerkin schemes based on continuous interpolation, and we refer to [72, 156, 155] for a review. The idea is to use the following approximate quadrature

$$\int_K \varphi_i u_h = \sum_{j \in K} \omega_j |K| \varphi_i(\bar{x}_j) u_j = \omega_i |K| u_i$$

This actually constrains the type of elements that one can consider, such that the quadrature formula associated to the integral of the basis functions is accurate enough. We refer to [72, 156, 155] for a detailed discussion.

As already discussed in section §4.4.1, Galerkin mass lumping in the standard  $P^1$  case is not polluting the second order of accuracy in space. This allows to lump one, or both the Galerkin integrals in (4.72), leading either to the formulation called *selective lumping* (SL) in [RA10]

$$|C_i| \frac{u_i^{k+1} - u_i^n}{\Delta t} - \int_{\Omega_h} \varphi_i \frac{\bar{\Delta} u^{k+1}}{\Delta t} = - \sum_{K \in K_i} \int_K \omega_i^K \bar{r}_h^{\text{RK}(k)} \quad (4.73)$$

or the formulation called *global lumping* (GL) in [RA10]

$$|C_i| \left( \frac{u_i^{k+1} - u_i^n}{\Delta t} - \frac{\bar{\Delta} u_i^{k+1}}{\Delta t} \right) = - \sum_{K \in K_i} \int_K \omega_i^K \bar{r}_h^{\text{RK}(k)} \quad (4.74)$$

#### 4.4.5 Fluctuations and signals ....

When  $\omega_i^K$  is specified to give back a RD scheme this allows a nicer interpretation. we can recast a RK-RD scheme using *e.g.* formulation F1 and global lumping as

**RK2 scheme :**

Step 1 :

$$|C_i| \frac{u_i^1 - u_i^n}{\Delta t} + \sum_{K \in K_i} \beta_i^K \int_K \nabla \cdot \mathcal{F}_h(u_h^n) = 0$$

Step 2 :

$$|C_i| \frac{u_i^{n+1} - u_i^1}{\Delta t} + \sum_{K \in K_i} \beta_i^K \int_K \left( \frac{u_h^1 - u_h^n}{\Delta t} + \nabla \cdot \frac{\mathcal{F}_h(u_h^n) + \mathcal{F}_h(u_h^1)}{2} \right) = 0$$

**RK3 scheme :**

Step 1 :

$$|C_i| \frac{u_i^1 - u_i^n}{\Delta t} + \sum_{K \in K_i} \beta_i^K \int_K \nabla \cdot \mathcal{F}_h(u_h^n) = 0$$

Step 2 :

$$|C_i| \frac{u_i^2 - (u_i^1 + u_i^n)/2}{\Delta t} + \sum_{K \in K_i} \beta_i^K \int_K \left( \frac{u_h^1 - u_h^n}{2\Delta t} + \nabla \cdot \frac{\mathcal{F}_h(u_h^n) + \mathcal{F}_h(u_h^1)}{4} \right) = 0$$

Step 3 :

$$|C_i| \frac{(u_i^{n+1} + u_i^n)/2 - u_i^2}{\Delta t} + \sum_{K \in K_i} \beta_i^K \int_K \left( 2 \frac{u_h^2 - u_h^n}{\Delta t} + \nabla \cdot \frac{\mathcal{F}_h(u_h^n) + 4\mathcal{F}_h(u_h^2) + \mathcal{F}_h(u_h^1)}{6} \right) = 0$$

which brings us back to the framework of P.L. Roe [218] : compute a local fluctuation, and correct nodal values using signals from surrounding elements.

#### 4.4.6 Results

The results reported here after are meant to give numerical evidence that the explicit schemes are indeed second order accurate, and that they provide results competitive with those of the implicit high order ones. In particular note that, in all computations we have set

$$\Delta t = \text{CFL} \min_{i \in \Omega_h} \frac{|C_i|}{\sum_{K \in K_i} \max_{j \in K} \rho(|\vec{a}_K \cdot \vec{n}_j|)} \quad (4.75)$$

with  $\text{CFL} \approx 1$ . This conditions corresponds to the global positivity of the LF scheme, and gives time steps of roughly half the size of the ones obtained with (4.38).

The first test is a grid convergence study on the Euler equations. The problem considered is the transport of a smooth vortex. We refer to [96] for a description. The test has been run with schemes based on formulation F1, using both selective and global lumping, and using both the RK2 and RK3 schemes. The spatial distribution schemes tested are the linear LDA and SUPG distribution (defined by (2.34) and denoted by SU in the figures), and the nonlinear blended LDAN and LLFs schemes.

The results are summarized on figure 4.21. All the schemes show second order of accuracy confirming our theoretical expectations. The LDAN provides the largest errors. However note that no effort has been made in adapting the definition of the blending parameter to this new formulation. We refer to [RA10] for more details.

We also report a comparison on the double Mach reflection problem already considered in section §4.3.1. On figure 4.22 we report the density contours relative to the results of the cell centered FV scheme and LN scheme already discussed in section §4.3.1, and those of the explicit LDAN and LLFs scheme with RK2 and global lumping.

The results are still very competitive with those of the FV scheme. The contact emanating from the triple point is thinner, and the resolution of the jet on the lower wall more resolved. The LN result is still better, but again we did not adapt the definition of the nonlinear schemes to this new framework, and applied blindly the same formulas used elsewhere, and the explicit schemes are about 10 times faster.

This is very promising.

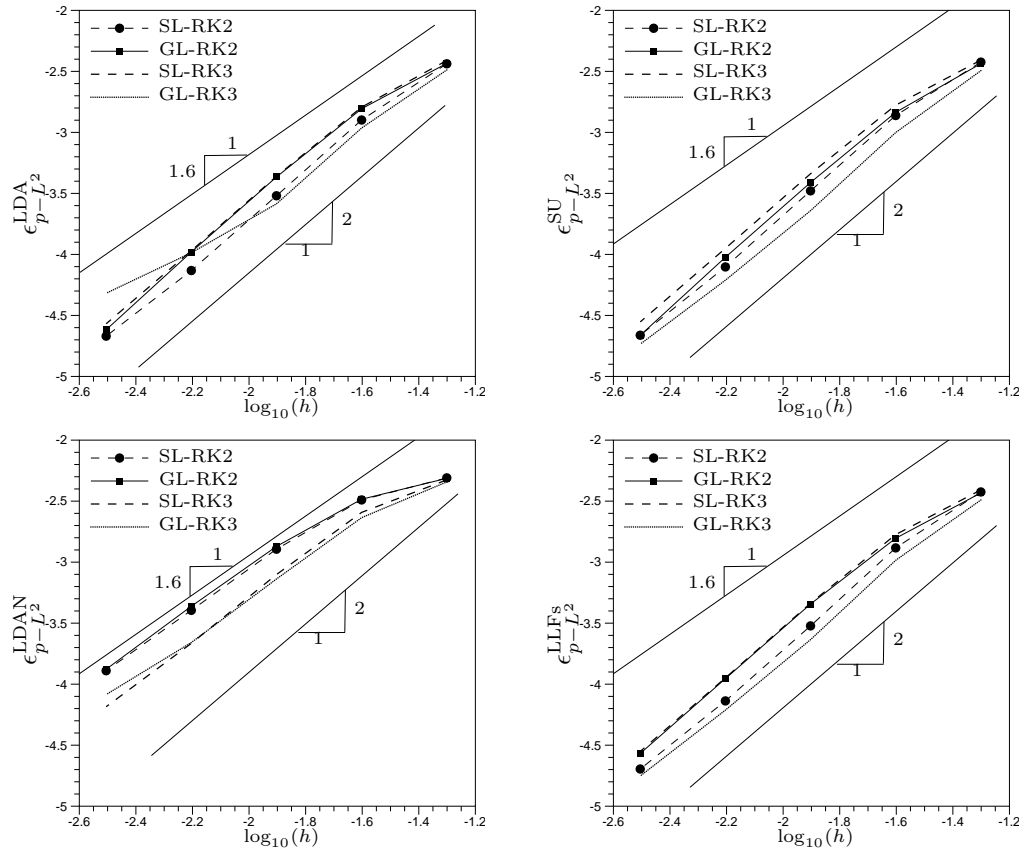


Figure 4.21: Vortex advection : grid convergence study.  $L^2$  pressure error for LDA (top-left), SU (top-right), LDAN (bottom-left) and LLFs (bottom-right) RK-RD schemes.

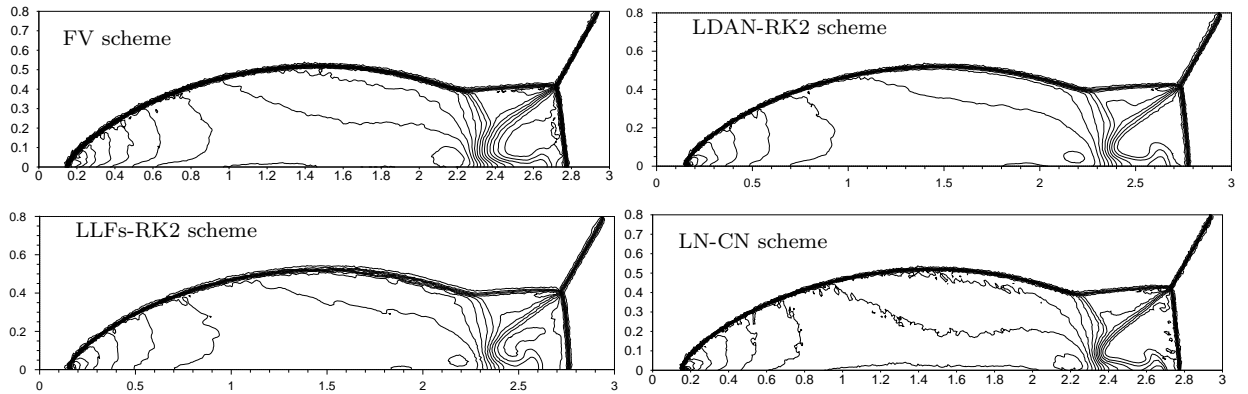


Figure 4.22: Double Mach reflection. Density at time  $t = 0.2$ . Cell-centered  $\mathcal{FV}$  scheme (top-left), LDAN-RK2 scheme (top-right), LLFs-RK2 scheme (bottom-left), and LN scheme (bottom-right)

## Part II

# Well-balanced discretizations for shallow water flows





After introducing some of the main issues related to the simulation of free surface flows using the Shallow Water model, this part will discuss the developments made in this field, using some of the RD schemes object of the first part.

Despite of the fact that the schemes proposed in this manuscript do not constitute the ultimate approach to shallow water simulation, some principles arise that could be used in other contexts as well. This is true especially for the construction of high order well balanced discretizations.

These contributions constitute

- an improved high order application of RD methods to free surface flow, including phenomena such as long wave run up, and tsunami simulation ; showing the full potential of the (second order) methods
- a *contribution to the understanding* of balanced discretizations for free surface flows, giving some *general principles* which can be used elsewhere
- an *important and promising alternative* to more classical approaches *to simulate real life free surface flows*



## Chapter 5

# Some numerical challenges in shallow water simulation

Free surface flows are relevant in a large number of applications, especially in civil and coastal engineering. The problems concerned are either (relatively) local, such as dam breaks and flooding, overland flows due to rainfall, nearshore wave propagation and interaction with complex bathymetries/structures, and tidal waves in rivers, or global such as in ocean or sea basin models for the study of *e.g.* tsunami generation and propagation.

The simulation of such flows can be carried by solving directly the three dimensional Navier-Stokes equations. However, for many applications, including *e.g.* nearshore wave propagation and flooding, simplified models obtained by combining vertical averaging and some form of thin layer approximation provide reliable results. The applicability of such models depends on the nature of the flow and on the hypotheses at their basis [165, 37].

The simplest among these models is the so-called Shallow Water model assuming that the waves that develop in the flow are *long* (small ratio amplitude/wavelength), and that the pressure has a hydrostatic vertical variation [117, 175].

The first order approximation (in terms of the ratio amplitude/wavelength) provides a non-homogeneous hyperbolic system where the effects of the variation of the bathymetry and the viscous friction on the bottom are modeled by source terms [117, 175]. More complex nonlinear models can be obtained by including higher order terms, and depending on the hypotheses on the flow [117, 175, 165, 37].

Part II of this manuscript, discusses my work on application and further development of residual distribution for the solution of the Shallow Water system that reads

$$\begin{aligned} \partial_t d + \nabla \cdot (d\vec{v}) + R(x, y, t) &= 0 \\ \partial_t (d\vec{v}) + \nabla \cdot (d\vec{v} \otimes \vec{v} + p(d)\mathbf{I}) + gd(\nabla b + k_f \vec{v}) &= 0 \end{aligned} \tag{5.1}$$

where  $d$  represents the depth,  $\vec{v}$  the (vertically averaged) local velocity,  $R$  is a source of mass (*e.g.* associated to rainfall),  $b$  is the bathymetry,  $k_f$  is a friction coefficient generally depending on the solution :

$$k_f = k_f(d, \vec{v}) \tag{5.2}$$

The hydrostatic pressure is given by

$$p(d) = g \frac{d^2}{2}$$

System (1.4) is endowed with a mathematical entropy coinciding with the total energy [245, 246, 130, 134], it is hyperbolic, and characterized by the physical constraint of the non-negativity of the depth. It is also useful to introduce the free surface level

$$\eta = d + b, \quad (5.3)$$

the *specific total energy*

$$\mathcal{E} = g\eta + k, \quad k = \frac{\|\vec{v}\|^2}{2}, \quad (5.4)$$

with  $k$  the kinetic energy, the discharge

$$\vec{q} = d\vec{v}, \quad (5.5)$$

and the Froude number

$$\text{Fr} = \frac{\|\vec{v}\|}{\sqrt{gd}} \quad (5.6)$$

playing for (5.1) the same role as the Mach number in gas dynamics.

The amount of literature related to the solution of (1.4) is extremely vast. This model finds applications in oceanography, hydrology, and meteorology (see *e.g.* [242, 142, 51, 112, 243, 244] and references therein). These applications may involve flows over irregular geometries, or water depths from the order of centimetres on very rough bed surfaces (flood propagation), or the run up on very complex structures (such as the bathymetry describing the ground elevation of a real coastal area). It is therefore necessary to develop accurate and stable discretizations to deal with these problems.

The main challenges when solving (5.1) numerically are mainly related to the discretization of the bathymetry and friction terms, and to the numerical treatment of nearly dry regions ( $d = 0$ ). For the first issue, one speaks often *asymptotic preserving* character or *well balancedness* of a discretization. The second issue is what is referred to as the wetting/drying strategy.

A discussion of these issues is given in the following sections, while in the following chapters we will discuss how these issues have been dealt with on unstructured grids using the residual approach discussed in part I.

## 5.1 Numerical challenges : well balancedness/C-property

When discretizing (5.1), one of the most important issues is to reproduce the balance between

- potential effects related to gravity, represented by the potential  $g\eta$  ;
- kinetic effects, related to the kinetic energy  $k$  ;
- dissipative effects represented by the friction  $k_f \vec{v}$

The balance between these three phenomena leads to a number physically relevant steady state equilibria. In order to be able to study perturbations of these states, and to avoid the pollution of numerical results by unwanted perturbations, one would like the numerical method to preserve as accurately as possible these steady equilibria. The typical example is the so called *lake at rest state* involving constant free surface  $\eta$ , that should be remaining flat whatever the shape of the bottom  $b(x, y)$ . This property is what one refers to as *Conservation property, or C-property*.

Starting from the initial work of [26] (see also [121]), the literature has been flooded by different ideas on how to handle the slope term  $g\nabla b$  in order to correctly preserve the lake at rest state. It is impossible to cite all of them, so the reader may refer to [26, 121, 111, 196, 276, 143, 89, 189, 19, 45] and references therein for an overview.

Suppose a scheme can be cast as

$$M \frac{dU}{dt} + R(U) = 0$$

with  $U$  the array containing all the discrete unknowns (nodal values, cell averages, polynomial coefficients etc etc.) on the mesh. The basic idea of all these methods is that, if  $U_0$  is the set of unknowns corresponding to a physically relevant steady equilibrium, then the numerical method should be desined or modified such that  $R(U_0) = 0$ .

Presently, practically all schemes can handle correctly the steady lake at rest state

$$d + b = \eta_0 = \text{const}, \quad \vec{v} = 0 \quad (5.7)$$

More recent developments are trying to handle more general states. An example is given by the one dimensional flow

$$g\eta + u^2/2 = \mathcal{E}_0 = \text{const}, \quad du = q_0 = \text{const}$$

There are several examples of methods that can handle these solutions exactly in one space dimension. The reader can refer to [197] for an overview. As in most methods allowing to preserve the lake at rest state, these approaches boil down to modifications of existing schemes that allow to express the discrete equations as differences and averages of the total energy, and the discharge, thus allowing to verify the condition  $R(U_0) = 0$ .

However, because based on a smart discrete differencing, these methods are inherently one dimensional. The consequence is that, while still working very well on structured cartesian meshes [196], they will perform as any other method on irregular grids.

A different steady equilibrium is obtained by including friction into the game. In one dimension, this state describes the steady flow in a sloping channel, with constant slope  $\partial_x b = -\xi_0 = \text{const}$ . In this case on can easily find the steady state  $d = d_0(\xi_0) = \text{const}$ ,  $u = u_0(\xi_0) = \text{const}$  which is determined by the conditions

$$du = q_0 = \text{const}, \quad k_f(d_0, u_0) u_0 = \xi_0$$

This state has some importance in hydrology (*e.g.* study of rainfall overflows). In this case, expressing the spatial derivatives in terms of the steady invariants  $d$  and  $u$  is not enough. The scheme must take explicitly into account the equilibrium between the different terms in order to preserve this state. The interested reader can consult the work of [59] and references

therein. In the paper, the authors show that finite volume fluxes need to be modified to include the effects of friction to preserve this state. This case is much less trivial. Friction not being a differential term, even powerful approaches as the differencing method based on singular residuals used in [197, 196] will not work. The modification of the scheme will depend on the definition of the FV flux.

## 5.2 Numerical challenges : wetting/drying

When approaching a dry region,  $d$  tends to zero and additional trouble has to be faced. Obviously, if  $d = 0$  everywhere, there is nothing to do, any method should reduce in this case to an identity  $0 = 0$ . The problem arises at a wet/dry interface, and when  $d \ll 1$  but  $d \neq 0$ . Techniques dealing with these issues are known as wetting/drying techniques.

Some problems are more related to implementation issues. For example, when using a conservative scheme, one solves for  $d$  and  $d\vec{v}$ , so for  $d \ll 1$  one is facing the question of how to get  $\vec{v}$  avoiding division by zero, and avoiding ill posed cases where both  $d\vec{v}$  and  $d$  are small, but not of the same order of magnitude.

A more interesting issue is how to preserve the condition  $d \geq 0$  numerically. In high order discretizations based on the combination of a robust first order FV flux plus some high order polynomial approximation/reconstruction, the key is the modification of the limiter in regions where  $d \ll 1$ . The principle is the same whether the high order approximation is based on a reconstruction [277], or on a in-cell polynomial representation [279, 106], or on a combination of them [101]. However, the details are different in each case.

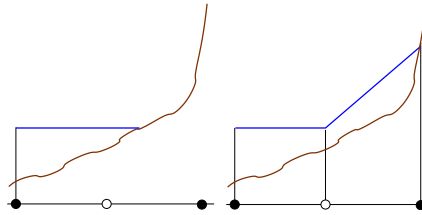


Figure 5.1: Lake at rest and dry areas. Left : physical state. Right : linear representation

Lastly, one must also make sure that if a physical steady equilibrium involving a dry state is present, the scheme should not destroy it. A typical example of how this could happen is shown in figure 5.1. In the figure, the left picture shows a physical steady state with water touching the ground. Suppose now to use a method that in the cell defined by the two black nodes used a sub-cell linear approximation. In this case, the linear reconstruction of the free surface, shown on the right picture, would induce a hydrostatic head which would push the flow toward the left.

The cure of this type of phenomenon has made the object of several papers, many of which from the spanish group revolving around M.E. Vasquez-Cendon and P.G. Navarro [47, 112, 46] (see however also [57, 58]). The idea is that close to the dry interface the *numerical slope* should be modified such that *if  $\eta = \text{const}$  and  $\vec{v} = 0$ , then the steady state is not perturbed*. In our example of figure 5.1, this would boil down to reduce the  $\Delta b$  used in the scheme in the right sub-cell, in order to satisfy  $\Delta\eta = 0$ .

# Chapter 6

## Contributions : well-balancing via residual schemes

### 6.1 Preliminaries

We consider the issue of well balancing in the context of residual distribution. It starts with the case of the lake at rest solution, which, for a particular implementation of the LDA and PSI schemes (cf. section §2.2.5), was analyzed for the first time in [45].

We then consider some generalizations. These generalizations, are based on the idea that all steady exact solutions are described by a certain operator being zero. This operator is not, in general, the conservative form of (5.1).

Let us start from the compact form of (5.1)

$$\partial_t u + \nabla \cdot \mathcal{F}(u) + \mathcal{S}(u, \nabla b) = 0 \quad \text{in } \Omega \times [0, T_{\text{fin}}] \subset \mathbb{R}^2 \times \mathbb{R}_+ \quad (6.1)$$

where in general

$$\mathcal{S} = gd \left[ 0, \nabla b + c_f \vec{v} \right]$$

#### 6.1.1 Super consistency analysis

We shall assume everywhere that *both the flux  $\mathcal{F}$  and the source term  $\mathcal{S}$  are at least Lipschitz continuous* :

$$\|\mathcal{F}(u) - \mathcal{F}(v)\| \leq K_{\mathcal{F}} \|u - v\| \quad \text{and} \quad \|\mathcal{S}(u, \nabla b) - \mathcal{S}(v, \nabla b)\| \leq K_{\mathcal{S}} \|u - v\| \quad (6.2)$$

Let us consider a set of derived variables  $v$  that depend on  $u$ , and might depend on  $b$  :  $v = v(u, b)$ . Examples are

- Total energy variables :  $v = [\mathcal{E}, \vec{q}]^t$
- Symmetrizing variables [130] :  $v = [g\eta - k, \vec{v}]^t$

Clearly, the flux not depending directly on  $b$  but only on  $u$ , the application of the chain rule leads to

$$\nabla \cdot \mathcal{F}(u(v, b)) = \vec{a}_v(u) \cdot \nabla v + \mathcal{S}_v(u, \nabla b) \quad (6.3)$$



where  $\mathcal{S}_v(u, \nabla b)$  is the additional contribution of all the terms containing derivatives of the bathymetry. Using this notation, we can prove the following result.

**Lemma 6.1.1** (Super consistency - local estimate). *Given an analytical bathymetry  $b$ , let  $v(u, b)$  be a set of invariants such that a family of steady equilibria for (5.1)-(6.1) is completely described by*

$$v = v_0 = \text{const}$$

Let  $\mathcal{F}_h = \mathcal{F}(u(v_h, b))$  and  $\mathcal{S}_h = \mathcal{S}(u(v_h, b), \nabla b)$ , with  $v_h$  the piecewise linear continuous  $P^1$  approximation of  $v$ . Assume that  $(v, b) \mapsto u$  is a one to one smooth mapping  $C^l$  with  $l$  sufficiently large, and similarly  $(v, b) \mapsto \mathcal{F}(u(v, b))$  is also  $C^{l'}$  with  $l'$  sufficiently large. Then, for exact integration we have

$$\phi^K(v_0, b) = \oint_{\partial K} \mathcal{F}_h \cdot \vec{n} + \int_K \mathcal{S}_h = 0$$

For approximate integration, let

$$\phi^K(v_0, b) = \sum_{f \in \partial K} \sum_{q=1}^{f_q} \omega_q \mathcal{F}_h(\vec{x}_q) \cdot \vec{n}_f + \sum_{q=1}^{v_q} \bar{\omega}_q \mathcal{S}_h(\vec{x}_q)$$

and let the line quadrature formula be exact for polynomials of degree  $p_f \geq 1$ , and the volume quadrature formula be exact for polynomials of degree  $p_v \geq 1$ . If  $b \in H^{p+1}(\Omega)$  with  $\nabla b \in H^p(\Omega_h)$  and with  $p \geq \min(p_f, p_v + 1)$ , then

$$|\phi^K(v_0, b)| \leq C h^r \quad \text{with} \quad r = \min(p_f + 2, p_v + 3)$$

*Proof.* For  $b \in H^{p+1}$  with  $p \geq \min(p_f, p_v + 1) \geq 1$  we can write for exact integration

$$\phi^K(v_0, b) = \oint_{\partial K} \mathcal{F}(u(v_h, b)) \cdot \vec{n} + \int_K \mathcal{S}(u(v_h, b), \nabla b) = \int_K \left( \vec{a}_v(v_0) \cdot \nabla v_0 + \mathcal{S}_v(v_0, \cdot \nabla b) + \mathcal{S}(v_0, b, \nabla b) \right)$$

Since  $v_0$  is an invariant describing a steady equilibrium, then from (6.3) and (6.1) we have

$$\nabla v_0 = 0, \quad \mathcal{S}_v(v_0, \cdot \nabla b) + \mathcal{S}(v_0, b, \nabla b) = 0 \quad \text{a.e.}$$

As a consequence, we deduce that  $\phi^K(v_0, b) = 0$ .

The second part of the proof uses the smoothness of the application  $(v, b) \mapsto \mathcal{F}$ . Due to the assumed regularity of  $(v, b) \mapsto \mathcal{F}(u(v, b))$ , and since  $v_h = v_0$  which is constant, then  $\mathcal{F}_h = \mathcal{F}(u(v_0, b))$  has the same regularity of  $b$ , which means that  $\mathcal{F}(u(v_0, b)) \in H^{p+1}(\Omega_h)$ . Similarly, we can argue that  $\mathcal{S}_h = \mathcal{S}(u(v_0, b), \nabla b)$  is in  $H^p$ .

Consider on each  $K$ , the polynomials  $\widehat{\mathcal{F}}_h$  of degree  $p_f$ , and the polynomials  $\widetilde{\mathcal{S}}_h$  of degree  $p_v$  such that (for simplicity we omit the additional superscript  $K$ )

$$\sum_{q=1}^{f_q} \omega_q \mathcal{F}_h(\vec{x}_q) \cdot \vec{n}_f = \int_f \widehat{\mathcal{F}}_h \cdot \vec{n}_f \quad \text{and} \quad \sum_{q=1}^{v_q} \bar{\omega}_q \mathcal{S}_h(\vec{x}_q) = \int_K \widetilde{\mathcal{S}}_h$$

With this notation, we can write, subtracting the exact integral which is zero :

$$\begin{aligned}
|\phi^K(v_0, b)| &= \left| \sum_{f \in \partial K} \int_f \widehat{\mathcal{F}}_h \cdot \vec{n}_f + \int_K \widetilde{\mathcal{S}}_h \right| \\
&= \left| \sum_{f \in \partial K} \int_f (\widehat{\mathcal{F}}_h - \mathcal{F}(u(v_0, b))) \cdot \vec{n}_f + \int_K (\widetilde{\mathcal{S}}_h - \mathcal{S}(u(v_0, b), \nabla b)) \right| \\
&\leq \sum_{f \in \partial K} \int_f |(\widehat{\mathcal{F}}_h - \mathcal{F}(u(v_0, b))) \cdot \vec{n}_f| + \int_K |\widetilde{\mathcal{S}}_h - \mathcal{S}(u(v_0, b), \nabla b)|
\end{aligned}$$

For the given regularity of  $b$ , we can write using standard approximation arguments [63, 105]

$$\begin{aligned}
|\widehat{\mathcal{F}}_h - \mathcal{F}(u(v_0, b))| &\leq C(v_0, b) h^{p_f+1} \Rightarrow \int_f |(\widehat{\mathcal{F}}_h - \mathcal{F}(u(v_0, b))) \cdot \vec{n}_f| = \mathcal{O}(h^{p_f+2}) \\
|\widetilde{\mathcal{S}}_h - \mathcal{S}(u(v_0, b), \nabla b)| &\leq C'(v_0, b) h^{p_v+1} \Rightarrow \int_K |\widetilde{\mathcal{S}}_h - \mathcal{S}(u(v_0, b), \nabla b)| = \mathcal{O}(h^{p_v+3})
\end{aligned}$$

This leads to the final estimate  $|\phi^K(v_0, b)| \leq C'' \max(h^{p_f+2}, h^{p_v+3})$ .  $\square$

This lemma allows to prove a more general result. In particular, we set

$$\phi^K(u_h^n) = \oint_{\partial K} \mathcal{F}_h^n \cdot \vec{n} + \int_K \mathcal{S}_h(u_h^n, \nabla b_h)$$

and we consider the two following two schemes :

1. Linearity preserving RD based on multistep time integration (cf. section §4.1.2, equations (4.4)-(4.6)-(4.7)) :

$$\begin{aligned}
\sum_{K \in \Omega_h} \beta_i^K \Phi^K(u_h) &= 0, \quad \forall i \in \Omega_h \\
\Phi^K(u_h) &= \sum_{i=0}^p \alpha_i \int_K \frac{\delta u_h^{n+1-i}}{\Delta t} + \sum_{j=0}^q \theta_j \phi^K(u_h^{n+1-j})
\end{aligned} \tag{6.4}$$

In particular, given an arbitrary compactly supported function  $\psi \in C_0^{l+1}(\Omega)$ , and replacing  $u_h$  by  $u(v_h, b) = u(v_0, b)$ , the approximation of a steady equilibrium corresponding to the analytical bathymetry  $b$ , and setting  $\mathcal{F}_h = \mathcal{F}(u(v_0, b))$  and  $\mathcal{S}_h = \mathcal{S}(u(v_0, b), \nabla b)$ , we can associate to this scheme the global space-time truncation error (cf. section §4.1.2, and lemma 6.1.1)

$$\epsilon(v_0, b, \psi) = \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{K \in \Omega_h} \sum_{j \in K} \psi_j \beta_j^K \phi^K(v_0, b) \tag{6.5}$$

2. Linearity preserving RK-RD scheme based on Runge-Kutta 2 time integration and mass lumping (cf. section §4.4.5) :

$$\begin{aligned} |C_i| \frac{u_i^1 - u_i^n}{\Delta t} &= - \sum_{K \in \Omega_h} \beta_i^K \phi^K(u_h^n) = 0, \quad \forall i \in \Omega_h \\ |C_i| \frac{u_i^{n+1} - u_i^1}{\Delta t} &= - \sum_{K \in \Omega_h} \beta_i^K \Phi^K(u_h) = 0, \quad \forall i \in \Omega_h \end{aligned} \quad (6.6)$$

where now

$$\Phi^K(u_h) = \int_K \frac{u_h^1 - u_h^n}{\Delta t} + \frac{1}{2} \phi^K(u_h^n) + \frac{1}{2} \phi^K(u_h^1)$$

In this case, given an arbitrary compactly supported function  $\psi \in C_0^{l+1}(\Omega)$ , and replacing  $u_h$  by  $u(v_h, b) = u(v_0, b)$ , the approximation of a steady equilibrium corresponding to the analytical bathymetry  $b$ , and setting  $\mathcal{F}_h = \mathcal{F}(u(v_0, b))$  and  $\mathcal{S}_h = \mathcal{S}(u(v_0, b), \nabla b)$ , we can associate to this scheme the global space-time truncation error (cf. section §4.1.2 and lemma 6.1.1)

$$\begin{aligned} \epsilon(v_0, b, \psi) &= \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{K \in \Omega_h} \sum_{j \in K} \psi_j \beta_j^K \phi^K(v_0, b) \\ &+ \frac{1}{2} \sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{K \in \Omega_h} \sum_{j \in K} \psi_j (\beta_j^K \phi^K(u_h^1) - \beta_j^K \phi^K(v_0, b)) \end{aligned} \quad (6.7)$$

The notation introduced allows to prove the following global result.

**Proposition 6.1.2** (Super consistency). *Under the hypotheses and with the notation of lemma 6.1.1, under the standard regularity assumptions on the mesh and on the time step*

$$C_1^h \leq \max_{K \in \Omega_h} \frac{h^2}{|K|} \leq C_2^h \quad \text{and} \quad C_1^{\Delta t} \leq \frac{\Delta t}{h} \leq C_2^{\Delta t}$$

and provided that the distribution coefficients  $\beta_i^K$  associated to the schemes are uniformly bounded w.r.t.  $h$ ,  $u_h$ , element residuals, and w.r.t. to the data of the problem, the linearity preserving RD schemes based on multistep time integration (6.4), and the explicit RK-RD based on Runge-Kutta 2 time stepping and mass lumping (6.6) preserve exactly the initial steady equilibrium for exact integration. For approximate integration, under the same hypotheses of lemma 6.1.1 they verify a global truncation error estimate of the type

$$|\epsilon(v_0, b, \psi)| \leq C h^l, \quad l = \min(p_f + 1, p_v + 2) \quad (6.8)$$

with the error  $\epsilon(v_0, b, \psi)$  given by (6.5) and (6.7), respectively.

*Proof.* For exact integration, lemma 6.1.1 guarantees that  $\phi^K(v_0, b) = 0$ . As a consequence, both families of schemes admit the trivial solution  $u_i^{n+1} = u(v_0, b_i) \forall i \in \Omega_h$  and  $\forall n \geq 0$ .

The second part of the proof is more involved. Following the the proof of lemma 6.1.1, we first introduce on each  $K$  the polynomials  $\widehat{\mathcal{F}}_h(u_h)$ , of degree  $p_f$ , and  $\widehat{\mathcal{S}}_h(u_h, \nabla b)$ , of

degree  $p_v$ , such that the numerical quadrature is equivalent to exact quadrature w.r.t these polynomials. Note that the dependence on  $u_h$  has been added for the following analysis. Next, we consider the Galerkin residuals

$$\phi_i^G(u_h) = \int_K \varphi_i (\nabla \cdot \widehat{\mathcal{F}}_h(u_h) + \widetilde{\mathcal{S}}_h(u_h, \nabla b))$$

which are assume to be evaluated *exactly*. This guarantees that (cf. lemma 6.1.1)

$$\sum_{j \in K} \phi_j^G(u_h) = \int_K (\nabla \cdot \widehat{\mathcal{F}}_h(u_h) + \widetilde{\mathcal{S}}_h(u_h, \nabla b)) = \sum_{j \in K} \beta_j^K \phi^K(u_h)$$

However, the exact evaluation also implies that each  $\phi_i^G(u_h)$  is computed by means of quadrature formulae one degree more accurate than those used to compute  $\phi^K(u_h)$ . In particular, it implies that the quadrature is exact for both the  $p_f$  degree polynomial  $\varphi_i \nabla \cdot \widehat{\mathcal{F}}_h$ , and for the  $p_v + 1$  degree polynomial  $\varphi_i \widetilde{\mathcal{S}}_h$ . As a consequence, with arguments similar to those used in the proof of lemma 6.1.1, and using the regularity assumptions on  $\psi$  we can state that

$$\begin{aligned} \sum_{j \in K} \psi_j \int_K \varphi_j \nabla \cdot \widehat{\mathcal{F}}_h(u_h) &= \int_K \psi_h \nabla \cdot \widehat{\mathcal{F}}_h(u_h) = \int_K \psi \nabla \cdot \mathcal{F}(u) + \mathcal{O}(h^{p_f+3}) \\ \sum_{j \in K} \psi_j \int_K \varphi_j \widetilde{\mathcal{S}}_h(u_h, \nabla b) &= \int_K \psi_h \widetilde{\mathcal{S}}_h(u_h, \nabla b) = \int_K \psi \mathcal{S}(u) + \mathcal{O}(h^{p_v+4}) \end{aligned} \quad (6.9)$$

We estimate now the error of the schemes based on multistep time integration. With the notation introduced, error (6.5) can be now recast as

$$\begin{aligned} \epsilon(v_0, b, \psi) &= \overbrace{\int_{t=0}^{t=T_{\text{fin}}} \int_{\Omega_h} \left( -\widehat{\mathcal{F}}_h(u(v_0, b)) \cdot \nabla \psi_h + \psi_h \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b) \right)}^{\text{I}} \\ &+ \underbrace{\sum_{n=0}^N \int_{t^n}^{t^{n+1}} \sum_{K \in \Omega_h} \frac{1}{3} \sum_{i,j \in K} (\psi_j - \psi_i) \left( \beta_j^K \phi^K(v_0, b) - \phi_j^G(u(v_0, b)) \right)}_{\text{II}} \end{aligned} \quad (6.10)$$

Proceeding as in the proof of lemma 6.1.1, we can easily argue that

$$\begin{aligned} |\phi_i^G(u_h)| &\leq C' h^2 \max(h^{-1} \|\widehat{\mathcal{F}}_h(u(v_0, b)) - \mathcal{F}(u(v_0, b))\|, \|\widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b) - \mathcal{S}(u(v_0, b), \nabla b)\|) \\ &\leq C(K) \max(h^{p_f+2}, h^{p_v+3}) \end{aligned}$$

We can thus use the result of lemma 6.1.1, the regularity of  $\varphi$ , and the uniform boundedness of the distribution coefficients  $\beta_i^K$ , to estimate term II as

$$|\text{II}| \leq C'_1(\Omega_h, T_{\text{fin}}), \Delta t^{-1} \Delta t h^{-2} \|\nabla \psi\|_{L^\infty(\Omega)} h \max(h^{p_f+2}, h^{p_v+3}) \leq C_1(\Omega_h, T_{\text{fin}}) h^l$$

with  $l$  as in (6.8) It remains to estimate term I. This is done first noting that, using the fact that  $u(v_0, b)$  is an exact steady solution

$$\begin{aligned} \text{I} &= \int_{t=0}^{t=T_{\text{fin}}} \int_{\Omega_h} \left( \mathcal{F}(u(v_0, b)) \cdot \nabla \psi - \widehat{\mathcal{F}}_h(u(v_0, b)) \cdot \nabla \psi_h \right) \\ &+ \int_{t=0}^{t=T_{\text{fin}}} \int_{\Omega_h} (\psi_h \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b) - \psi \mathcal{S}(u(v_0, b), \nabla b)) \end{aligned}$$

Using now (6.9), and the fact that the number of triangles in a two dimensional grid is of  $\mathcal{O}(h^{-2})$ , we immediately deduce

$$|\text{I}| \leq C(\Omega_h, T_{\text{fin}}) h^l$$

with  $l$  as in (6.8). This achieves the proof for the multistep time integration case.

For the explicit schemes, we start noting that the error (6.7) can be easily recast as (cf. equation (6.10))

$$\begin{aligned} \epsilon(v_0, b, \psi) &= \text{I} + \text{II} \\ &+ \frac{\Delta t}{2} \sum_{n=0}^N \left\{ \overbrace{\int_{\Omega_h} \left( -(\widehat{\mathcal{F}}_h^1 - \widehat{\mathcal{F}}_h(u(v_0, b))) \cdot \nabla \psi_h + \psi_h (\widetilde{\mathcal{S}}_h^1 - \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b)) \right)}^{\text{III}} \right. \\ &+ \underbrace{\sum_{K \in \Omega_h} \sum_{i, j \in K} \frac{\psi_j - \psi_i}{3} \left( \beta_j^K(u_h^1) \phi^K(u_h^1) - \beta_j^K \phi^K(v_0, b) \right.} \\ &\quad \left. \left. - (\phi_j^G(u_h^1) - \phi_j^G(u(v_0, b))) \right) \right\} \end{aligned} \quad (6.11)$$

where the dependence of the distribution coefficients on the solution has been added, with I and II as in (6.10), and where  $\widehat{\mathcal{F}}_h^1 = \mathcal{F}(u(v_h^1, b))$  and  $\widetilde{\mathcal{S}}_h^1 = \mathcal{S}(u(v_h^1, b), \nabla b)$ , with  $v_h^1$  the approximation obtained from the nodal values  $v_i(u_i^1, b)$ , with  $u_i^1$  obtained from the first in (6.6) when  $u_h^n = u(v_0, b)$ . To end the proof we need to estimate III and IV.

To to this, we use the Lipschitz continuity of the flux to write

$$\|\widehat{\mathcal{F}}_h(u(v_h^1, b)) - \widehat{\mathcal{F}}_h(u(v_0, b))\| \leq K_{\mathcal{F}} \|u^1(v_h^1, b) - u(v_0, b)\|$$

The first in (6.6), lemma 6.1.1, and the regularity of the time stepping lead to

$$\|\widehat{\mathcal{F}}_h(u(v_h^1, b)) - \widehat{\mathcal{F}}_h(u(v_0, b))\| \leq \mathcal{O}(h^l)$$

with  $l$  as in (6.8). Similarly, the regularity of  $b$  and (6.2) allow to write

$$\|\widetilde{\mathcal{S}}_h(u(v_h^1, b), \nabla b) - \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b)\| \leq K_{\mathcal{S}} \|u^1(v_h^1, b) - u(v_0, b)\| \leq \mathcal{O}(h^l)$$

The last two estimates and standard approximation properties lead immediately to

$$\left| \frac{\Delta t}{2} \sum_{n=0}^N \text{III} \right| \leq \overline{C}'(\Omega_h, T_{\text{fin}}) \Delta t \Delta t^{-1} \|\nabla \psi\|_{L^\infty(\Omega)} \max(h^{p_f+1}, h^{p_v+2}) \leq \overline{C}(\Omega_h, T_{\text{fin}}) h^l$$

with  $l$  as in (6.8). Finally, using the same arguments, we can deduce easily that

$$\begin{aligned} \|\phi^K(u_h^1) - \phi^K(v_0, b)\| &= \left\| \oint_{\partial K} (\widehat{\mathcal{F}}_h(u(v_h^1, b)) - \widehat{\mathcal{F}}_h(u(v_0, b))) \cdot \vec{n} \right. \\ &\quad \left. + \int_K (\widetilde{\mathcal{S}}_h(u(v_h^1, b), \nabla b) - \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b)) \right\| \leq \mathcal{O}(h^{l+1}) \end{aligned}$$

and similarly

$$\begin{aligned} \|\phi_j^G(u_h^1) - \phi_j^G(u(v_0, b))\| &= \left\| \int_K \varphi_j \nabla \cdot (\widehat{\mathcal{F}}_h(u(v_h^1, b)) - \widehat{\mathcal{F}}_h(u(v_0, b))) \right. \\ &\quad \left. + \int_K \varphi_j (\widetilde{\mathcal{S}}_h(u(v_h^1, b), \nabla b) - \widetilde{\mathcal{S}}_h(u(v_0, b), \nabla b)) \right\| \leq \mathcal{O}(h^{l+1}) \end{aligned}$$

We can thus first estimate

$$\begin{aligned} \|\beta_j^K(u_h^1) \phi^K(u_h^1)\| &\leq \|\beta_j^K(u_h^1) \phi^K(v_0, b)\| + \|\beta_j^K(u_h^1) (\phi^K(u_h^1) - \phi^K(v_0, b))\| \\ &\leq \sup_{j, K \in \Omega_h} \|\beta_j^K(u_h^1)\| (C'_a h^{l+1} + C''_a h^{l+1}) \leq C_a h^{l+1} \end{aligned}$$

with  $l$  as in (6.8), and having used lemma 6.1.1. Term IV can thus be estimated, using the regularity of  $\psi$ , the fact that the number of time steps is of  $\mathcal{O}(\Delta t^{-1})$ , and that the number of triangles is of  $\mathcal{O}(h^{-2})$ , and the estimate on the Galerkin terms :

$$\left| \frac{\Delta t}{2} \sum_{n=0}^N \text{IV} \right| \leq \tilde{C}'(\Omega_h, \text{T}_{\text{fin}}) \Delta t \Delta t^{-1} h^{-2} \|\nabla \psi\|_{L^\infty(\Omega)} h (C_a + \sup_{j, K \in \Omega_h} \|\beta_j^K\| C'_a + C''_a) h^{l+1} \leq \tilde{C}(\Omega_h, \text{T}_{\text{fin}}) h^l$$

which achieves the proof.  $\square$

**Remark 6.1.3** (Bathymetry representation and regularity). *The last proposition shows that for finite time computations, if the bathymetry is regular enough, the solution convergence with a rate  $l > 2$ , as soon as the quadrature formulae are more than second order accurate. The proposition explicitly uses the assumption that an analytical bathymetry is used in the discretization, and that the regularity of this expression is such that the full accuracy of the quadrature formulas is recovered. In general, if the hypothesis  $p \geq \min(p_f, p_v + 1)$  is relaxed to  $p \geq 0$ , one can prove with the exact same arguments that*

$$\|\phi^K(v_0, b)\| = \mathcal{O}(h^r), \quad r = \min(p + 2, p_f + 2, p_v + 3)$$

and similarly that schemes (6.4) and (6.6) verify a truncation error estimate

$$\epsilon(v_0, b, \psi) = \mathcal{O}(h^l), \quad l = \min(p + 1, p_f + 1, p_v + 2) \quad (6.12)$$

**Remark 6.1.4** (Stability, consistency and convergence). *As already remarked in section §2.2.3 (cf. remark 2.2.10), the integral truncation error analysis used in this work only gives information about the consistency, and in the case of proposition 6.1.2 the super consistency, of the discretization. Without a stability criterion, there is a priori no guarantee that the rates of convergence of the proposition are actually observed in practice, since there is no guarantee of the existence of a unique discrete solution, and the presence of spurious modes reducing the rates of convergence, or even preventing it altogether, cannot be ruled out.*

Proposition 6.1.2 is the rationale for the developments made in this work. First results have been already published in [Ric11]. In the following sections, we will apply the proposition to some particular equilibria.

## 6.2 Basic C-property

The C-property is met if the lake at rest state (5.7) is preserved exactly. The first to analyze this solution in the RD context have been Brufau and Garcia-Navarro [45]. Their approach has been generalized in [RAD07, RB09b, Ric11]. These results are summarized in the following general proposition.

**Proposition 6.2.1** (C-property). *Let  $\omega_i(x, y, u_h)$  a uniformly bounded test function, such that  $\omega_i^K = \omega_i|_K$  is a polynomial of degree  $p_\omega$  in space. Any scheme that writes*

$$\sum_{K \in K_i} \int_K \omega_i \partial_t u_h + \sum_{K \in K_i} \oint_{\partial K} \omega_i \mathcal{F}(u_h) \cdot \vec{n} - \sum_{K \in K_i} \int_K \mathcal{F}(u_h) \cdot \nabla \omega_i + \sum_{K \in K_i} \int_K \omega_i \mathcal{S}_h(u_h, \nabla b) = 0$$

- Preserves exactly the lake at rest state, if  $d_h = \eta_h - b_h$ , and  $\mathcal{S}_h(u_h, \nabla b) = \mathcal{S}_h(u_h, \nabla b_h)$ , with  $b_h$  based on the same  $P^k$  Lagrange approximation used for  $\eta_h$ , and provided that all the integrals are evaluated exactly w.r.t. polynomials of degree  $p_\omega + 2k$  ;
- Is super-consistent with the lake at rest state, in the sense of proposition 6.1.2, if  $d_h = \eta_h - b$ , with  $b$  a smooth enough regular bathymetry. In particular, it verifies a local super consistency estimate

$$\phi_i^K = \sum_{K \in K_i} \oint_{\partial K} \omega_i \mathcal{F}(u_h) \cdot \vec{n} - \sum_{K \in K_i} \int_K \mathcal{F}(u_h) \cdot \nabla \omega_i + \sum_{K \in K_i} \int_K \omega_i \mathcal{S}_h(u_h, \nabla b) = \mathcal{O}(h^{l+1})$$

and truncation error estimate (6.8) when the numerical quadrature is exact for polynomials of degree  $p_\omega + p_f$  on the faces, and for polynomials of degree  $p_\omega + p_v$  on  $K$ , and with  $l$  as in (6.8).

*Proof.* The first part of the proof is immediately shown by considering that, because of the hypotheses made on  $d_h$ ,  $b_h$ , and on the integration formulas, and considering that on the lake at rest we have  $\eta_h = \eta_0 = \text{const}$  and  $\vec{v} = 0$ , we have

$$\begin{aligned} g \int_K \omega_i^K d_h \nabla b_h &= g \int_K \omega_i^K d_h \nabla (\eta_h - d_h) = -g \int_K \omega_i^K \nabla \frac{d_h^2}{2} \\ &= -g \oint_{\partial K} \omega_i^K \frac{d_h^2}{2} \vec{n} + g \int_{\partial K} \frac{d_h^2}{2} \nabla \omega_i^K = - \oint_{\partial K} \omega_i \mathcal{F}(u_h) \cdot \vec{n} + \sum_{K \in K_i} \int_K \mathcal{F}(u_h) \cdot \nabla \omega_i \end{aligned}$$

exactly. Hence, the source term integral balances *exactly* the flux integral.

The proof of the super consistency is achieved following the steps of the proof of lemma 6.1.1 and proposition 6.1.2. In particular, first we note that the mapping  $d_h = \eta_h - b$  is linear, and then that the quadrature is by hypothesis exact for  $\omega_i^K \widehat{\mathcal{F}}_h$  (face integrals),  $\widehat{\mathcal{F}}_h \cdot \nabla \omega_i^K$  and  $\omega_i^K \widetilde{\mathcal{S}}_h$  (volume integrals), where  $\widehat{\mathcal{F}}_h$  is a polynomial of degree  $p_f$  on  $K$  and  $\widetilde{\mathcal{S}}_h$  a polynomial

of degree  $p_v$ . Using this fact, we write

$$\begin{aligned} \oint_{\partial K} \omega_i \widehat{\mathcal{F}}_h \cdot \vec{n} - \int_K \widehat{\mathcal{F}}_h \cdot \nabla \omega_i + \int_K \omega_i \widetilde{\mathcal{S}}_h(u_h, \nabla b) &= \int_K \omega_i^K (\nabla \cdot \widehat{\mathcal{F}}_h + \widetilde{\mathcal{S}}_h(u_h, \nabla b)) \\ &= \int_K \omega_i^K (\nabla \cdot (\widehat{\mathcal{F}}_h - \mathcal{F}) + \widetilde{\mathcal{S}}_h(u_h, \nabla b) - \mathcal{S}) \end{aligned}$$

The next step is to introduce estimates for the approximation error  $\widehat{\mathcal{F}}_h - \mathcal{F}$  and  $\widetilde{\mathcal{S}}_h - \mathcal{S}$ , using standard approximation results [63, 105]. The rest follows exactly as in the proof of lemma 6.1.1 and of proposition 6.1.2 (omitted).  $\square$

**Remark 6.2.2** (Linearity preserving schemes). *Last proposition adds to the result of proposition 6.1.2 an exact preservation condition. In particular, when  $\omega_i^K$  is constant, the scheme reduces to a conservative linearity preserving scheme RD for which exact preservation is obtained for  $d_h = \eta_h - b_h$  and exact integration for polynomials of degree  $2k$  (on  $P^k$  Lagrange elements), giving back the results of [RAD07, RB09b].*

These properties have been numerically verified for linearity preserving RD schemes in [RAD07, RB09b, Ric11], to which we refer for a thorough discussion. As an example, we report on figure 6.1 the results obtained on the classical test proposed initially in [171], and involving a perturbation of the lake at rest state on a two-dimensional smooth bathymetry. The figure shows the results obtained with the LLFs-RK2 RD scheme on an unstructured triangulation in terms of report 3d views and line plots of the free surface (properly rescaled for the sake of plotting). As expected, the lake at rest is kept exactly away from the perturbation.

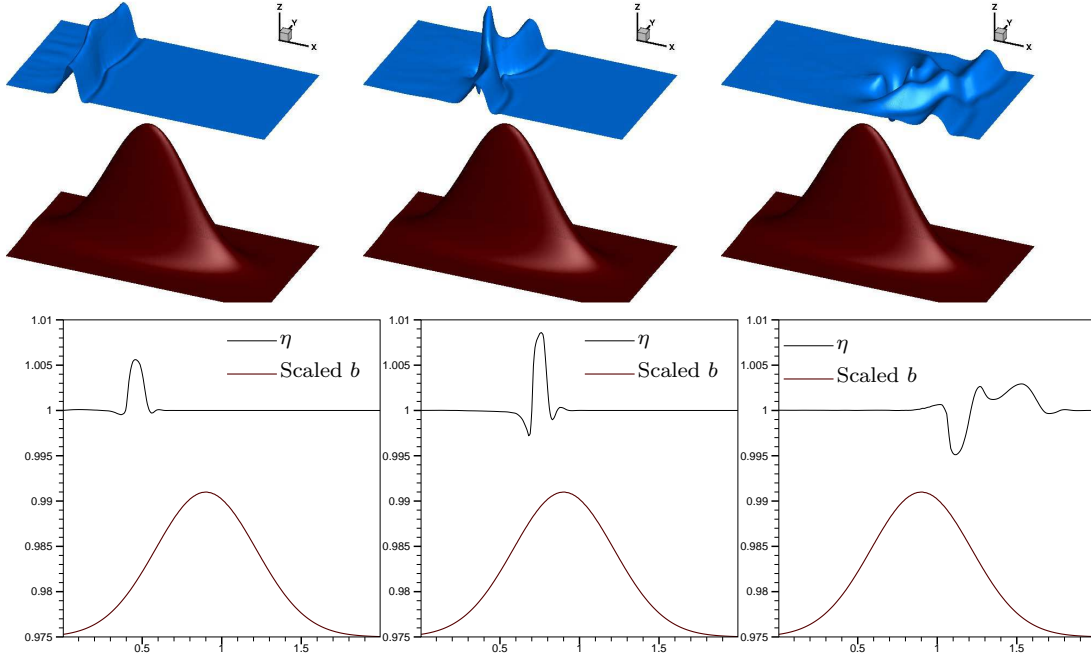


Figure 6.1: Perturbation of the lake at rest over a smooth bathymetry.



## 6.3 Generalizations of the C-property

### 6.3.1 Constant total energy flows

We start by considering, on two-dimensional unstructured grids, the approximation of pseudo one dimensional *frictionless homo-energetic pseudo-1d flows*. These flows are characterized by the invariance of the total energy and of the discharge, namely :

$$\begin{aligned}\mathcal{E} &= g\eta + \frac{\vec{v} \cdot \vec{v}}{2} = \mathcal{E}_0 = \text{const} \\ \vec{q} &= d\vec{v} = \vec{q}_0 = \text{const}\end{aligned}\tag{6.13}$$

By expressing all the spatial derivatives in terms of variations of these quantities we get

$$\begin{aligned}\partial d + \nabla \cdot \vec{q} &= 0 \\ \partial_t(d\vec{v}) + (\vec{v} \cdot \nabla) \vec{q} - (\vec{v}^\perp \cdot \nabla) \vec{q}^\perp \\ &+ \frac{1}{1 - \text{Fr}^2} \left( \frac{1}{g} (gd\nabla\mathcal{E} - \vec{v}\vec{v} \cdot \nabla\mathcal{E}) + \frac{\vec{v}}{gd} \vec{v} \cdot (\nabla\vec{q} \cdot \vec{v}) - \text{Fr}^2 (\nabla\vec{q})^t \cdot \vec{v} \right) = \frac{\vec{v}^\perp \cdot \nabla b}{1 - \text{Fr}^2} \vec{v}^\perp\end{aligned}\tag{6.14}$$

with  $\vec{v}^\perp = (-v_y, v_x)$ ,  $\vec{q}^\perp = d\vec{v}^\perp$ , and  $^t$  the transpose operator. Last equations show that indeed there exist an admissible family of steady solutions whose invariants are the total energy and the discharge. These solutions are constrained by the compatibility condition for the bathymetry  $\vec{v}^\perp \cdot \nabla b = 0$ , allowing bathymetry variations only in the direction of the discharge. This makes these solutions basically one-dimensional flows in the  $\vec{v}$  direction.

For a given value of  $b = b_0$ , and given the set of invariants  $v_0 = [\mathcal{E}_0, \vec{q}_0]^t$ ,  $d$  and  $\vec{v}$  are roots of a cubic polynomial that can be either written in terms of  $d$ , or in terms of the norm of the velocity (the direction being the same as  $\vec{q}_0$ )

$$\begin{cases} p(d) = d^3 - a_d d^2 + b_d \\ a_d = (\mathcal{E}_0 - gb_0)/g \\ b_d = \|\vec{q}_0\|^2/2 \end{cases}, \quad \begin{cases} p(u) = u^3 - a_u u^2 + b_u \\ a_u = 2(\mathcal{E}_0 - gb_0) \\ b_u = 2g\|\vec{q}_0\| \end{cases}\tag{6.15}$$

with the notation  $u = \|\vec{v}\|$ . It can be easily shown that

**Proposition 6.3.1** (Smoothness of total energy variables mapping). *Provided that*

$$\frac{2}{g^2} \frac{(\mathcal{E}_0 - gb_0)^3}{\|\vec{q}_0\|^2} > \frac{27}{4}$$

then

- $p_d(d)$  admits a unique solution  $d > 2a_d/(3g)$ , corresponding to a sub-critical flow ;
- $p_u(u)$  admits a unique solution  $u > \sqrt{a_u/3}$ , corresponding to a super-critical flow ;

In this case, the behavior of  $d$  and  $u$  is the same as  $b^\alpha$  with  $|\alpha| < 1$ .

We omit the proof, based on a (relatively simple) algebraic study of the polynomials. We use this family of solutions to illustrate numerically the consequences of proposition 6.1.2.

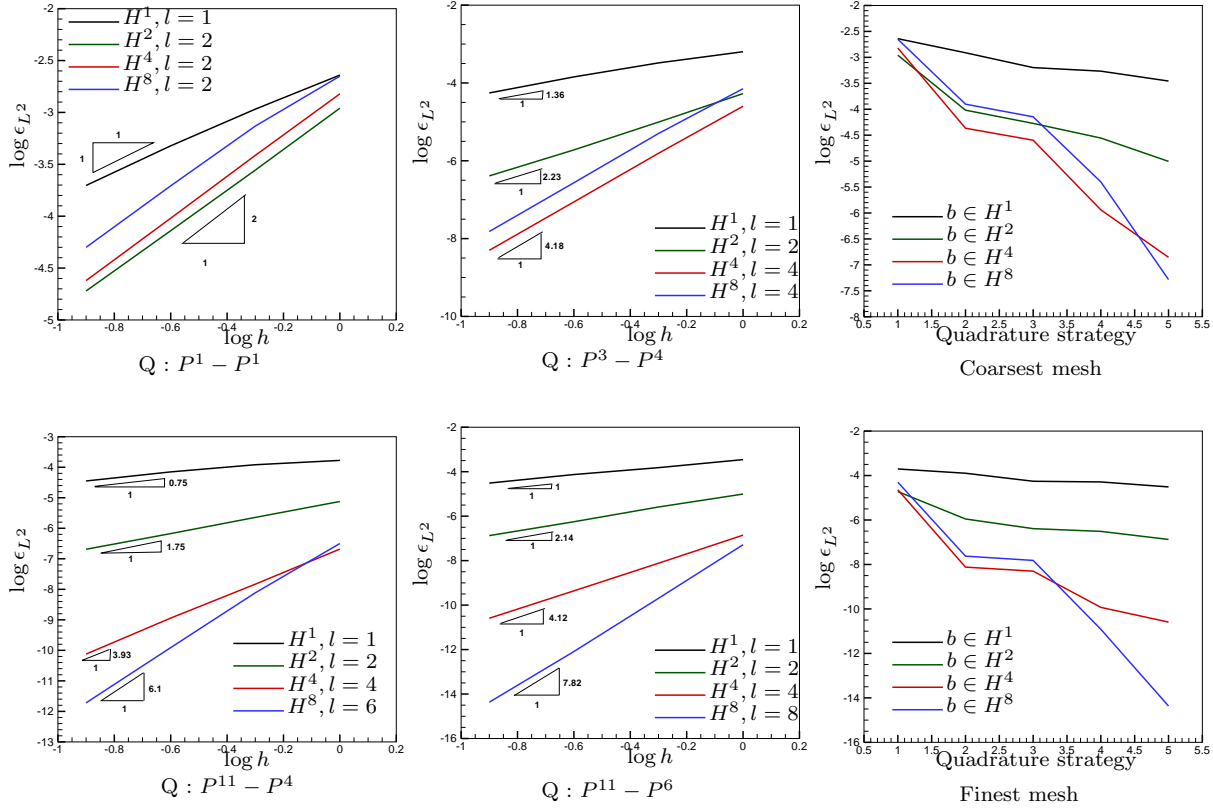


Figure 6.2: Super consistency with pseudo 1d flows : grid and quadrature convergence of the depth when changing the regularity of the bathymetry. Results of the LLFs-RK2 scheme

In particular, we repeat the test proposed in [Ric11]. On the square  $[0, 25]^2$ , we consider a bathymetry defined as

$$b(x, y) = \begin{cases} f(x) & \text{if } x \in [8, 12] \\ 0 & \text{otherwise} \end{cases}$$

The function  $f(x)$  is chosen to obtain increasing regularity. We start with the  $C^0$  definition  $f = 0.2 - (x - 10)^2/20$ , giving a  $H^1$  bathymetry, and consider  $C^p$  definitions involving even powers of the sin function, yielding  $H^{p+1}$  bathymetries. We compute initial nodal values from (6.13), with  $\mathcal{E}_0 = 22.06605$  and  $q_0 = (4.42, 0)$ , and run unsteady computations until time  $T_{\text{fin}} = 0.1$  on four nested unstructured grids (rightmost picture on figure 6.3).

In all the computations, we use the spatial approximation  $\mathcal{F}_h = \mathcal{F}(u(v_h, b))$ , with  $b$  the analytical bathymetry, and  $v_h$  linear (cf. proposition 6.1.2). The runs are repeated with different quadrature strategies. The results are summarized on figure 6.2. In the figure, the first four pictures on the left (first and second column) represent the grid convergence of the depth at fixed  $T_{\text{fin}}$  for different quadrature strategies, and different regularity of the bathymetry. Below each picture, we have reported the polynomials integrated exactly by the formulas used (the first corresponding to the face integrals, the second to element integrals).

In the figures, we also report in the legend the theoretical rate obtained from the more general estimate (6.12). In particular, for a  $H^{p+1}$  bathymetry, the accuracy measured for a finite time computation should be

$$\epsilon = \mathcal{O}(h^l) \text{ with } Q : P^f - P^v \implies l = \min(p + 1, f + 1, v + 2) \quad (6.16)$$

The scheme used is the LLFs-RK2 scheme, which is in general second order accurate.

We can see that the discrete solution at time  $T_{\text{fin}}$  super converges if the bathymetry is regular enough. The degree of super convergence depends also on the quadrature formula. The slope reduction is observed in all the cases in which  $b$  lacks enough continuity. We find the exact asymptotic behavior announced by proposition 6.3.1.

The last two pictures on figure 6.2, show the error convergence on a fixed grid (the coarsest on top, the finest on the bottom) when increasing the accuracy of the quadrature<sup>1</sup>. These plots confirm that the error indeed converges to zero (towards exact preservation) if the quadrature accuracy is increased. The smoother the bathymetry, the faster the convergence. Similar results were shown on [Ric11] with the LLFs-CN scheme for both solutions (6.13), and for the lake at rest case.

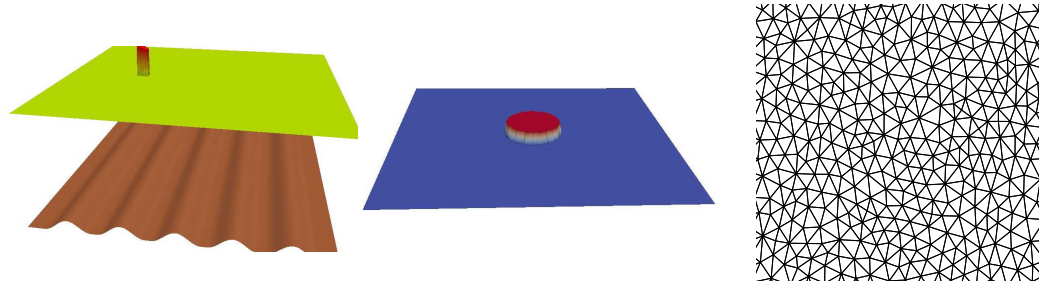


Figure 6.3: Left : Perturbation of pseudo 1d flow in a ribbed channel - total energy at time  $t = 0$ . Middle : Perturbation of channel flows with friction - depth at time  $t = 0$ . Right : triangular mesh used for the 2d simulations

**Remark 6.3.2** (On the expected rates). *The error dependence on the mesh size is not fully characterized by the analysis of proposition 6.1.2. The numerical results show that, for a fixed smoothness of the bathymetry, increasing the quadrature accuracy leads to a convergence of the solution. This dependence of the error on the choice of the formula is not captured by the analysis of proposition 6.1.2.*

To show *visually* the benefits of using our residual approach in conjunction with the approximation in total energy variables, we consider the perturbation of a pseudo one dimensional flow on a bathymetry representative of a ribbed channel. The bathymetry can be seen in the leftmost picture on figure 6.3. The initial solution is computed from (6.13) with a sub-critical initial state, and then a perturbation is added on the depth. The propagation of the perturbation is simulated on the mesh shown on the rightmost picture on figure 6.3.

<sup>1</sup>The scale of the x axis is not directly linked to the accuracy of the formula, which does however increase when running in the positive direction

The results obtained with the LLFs-RK2 scheme are reported on figure 6.4 where a 3d view of the temporal evolution of the energy is shown. The result shows the nice preservation of the initial state away from the perturbation, and its multidimensional evolution.

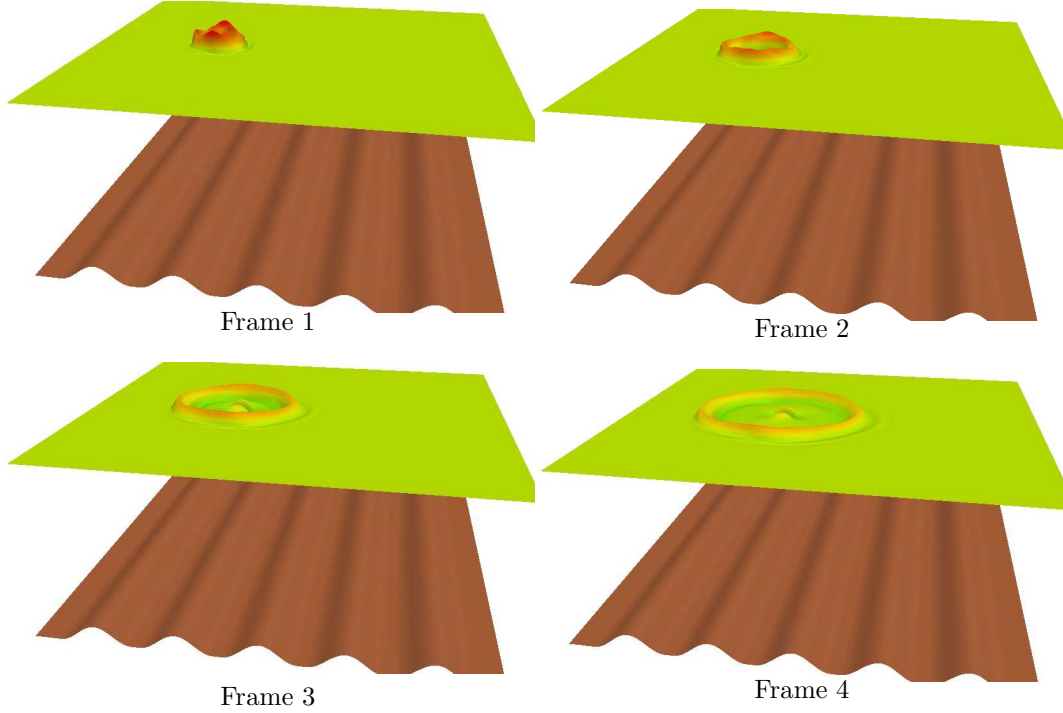


Figure 6.4: Perturbation of pseudo 1d flow in a ribbed channel : total energy. Solution of the LLFs-RK2 scheme

**Remark 6.3.3** (Bathymetry representation). *The results shown, as well as the theoretical developments, are based on the assumption that an analytical bathymetry is available, and that its exact form is used in the discretization. The availability of such a bathymetry is of course questionable. However, given the uncertainties in the experimental data providing such quantity, it seems possible to think of a pre-processing step building a spline (or some other smooth polynomial) representation of the data, thus giving some interest to the method.*

*Additionally, when dealing with irregular (or piecewise regular) bathymetries, the use of total energy variables interpolation is not justified, leading to first order of convergence, while the use of a  $C^0$  Lagrange approximation allows to recover second order rates [RB09b, RB09a]. However, an inspection of equations (6.14) suggests that the real information needed to correctly reproduce these solutions is  $\vec{v}^\perp \cdot \nabla b = 0$ , so that some a different approach based on the addition of a crosswind correction is perhaps possible.*

### 6.3.2 Flows in sloping channels with friction

We consider now the case in which the friction is included in the equations. The simplest equilibrium obtained in this case is characterized by a balance between bathymetry variations

(slope) and friction for a give flow rate, namely (cf. equation (5.1)) :

$$\begin{aligned} \vec{q} &= h\vec{v} = \vec{q}_0 = \text{const} \\ \nabla b + c_f(d, \vec{v})\vec{v} &= 0 \end{aligned}$$

These flows are again one-dimensional in the direction of the constant bathymetry gradient, which is, again, the direction of the velocity. They are characterized by constant depth, discharge and velocity, namely :

$$\begin{aligned} d &= d_0 = \frac{\|\vec{q}_0\|}{\|\vec{v}_0\|} \\ c_f\left(\frac{\|\vec{q}_0\|}{\|\vec{v}_0\|}, \vec{v}_0\right)\vec{v}_0 + \nabla b_0 &= 0 \end{aligned} \tag{6.17}$$

In particular, if the friction coefficient is given by Manning's law

$$c_f = \frac{n^2 \|\vec{v}\|}{d^{4/3}} \tag{6.18}$$

with  $n$  the Manning's coefficient, then the solution (6.17) gives

$$\begin{aligned} d &= d_0 = \left( \frac{n^2 \|\vec{q}_0\|^2}{\|\nabla b_0\|} \right)^{\frac{3}{10}} \\ \vec{v} = \vec{v}_0 &= - \left( \frac{\|\nabla b_0\| \|\vec{q}_0\|^{4/3}}{n^2} \right)^{\frac{3}{10}} \frac{\nabla b_0}{\|\nabla b_0\|} \end{aligned} \tag{6.19}$$

We have the following result.

**Proposition 6.3.4** (Preservation of sloping channels flows). *Let  $\omega_i(x, y, u_h)$  a uniformly bounded test function, such that  $\omega_i^K = \omega_i|_K$  is a polynomial of degree  $p_\omega$ . Any scheme that writes*

$$\sum_{K \in K_i} \int_K \omega_i \partial_t u_h + \sum_{K \in K_i} \oint_K \omega_i \mathcal{F}(u_h) \cdot \vec{n} - \sum_{K \in K_i} \int_K \mathcal{F}(u_h) \cdot \nabla \omega_i + \sum_{K \in K_i} \int_K \omega_i \mathcal{S}_h(u_h, \nabla b) = 0$$

with  $\mathcal{F}_h = \mathcal{F}(u_h)$ , preserves exactly sloping flows with friction.

*Proof.* In this case the flux is constant and gives no contributions to the equation, while by construction  $\mathcal{S} = gh_0[0, \nabla b_0 + c_f \vec{v}_0] = 0$ , so that all the integrals balance out whatever the (consistent) integration strategy.  $\square$

We provide two examples in which we perturb a steady solution. In particular, we consider two states that satisfy (6.19), one giving a supercritical flow, and the other a subcritical flow. We then perturb the depth, as shown on the middle picture on figure 6.3 and compute the evolution on the unstructured triangulation in the rightmost picture on the same figure.

The evolution of the perturbation computed by the LLFs-RK2 scheme is shown on figure 6.5. Again the results show the perfect preservation of the steady state on unstructured grids, and a nice and clean capturing of the evolution of the initial perturbation.

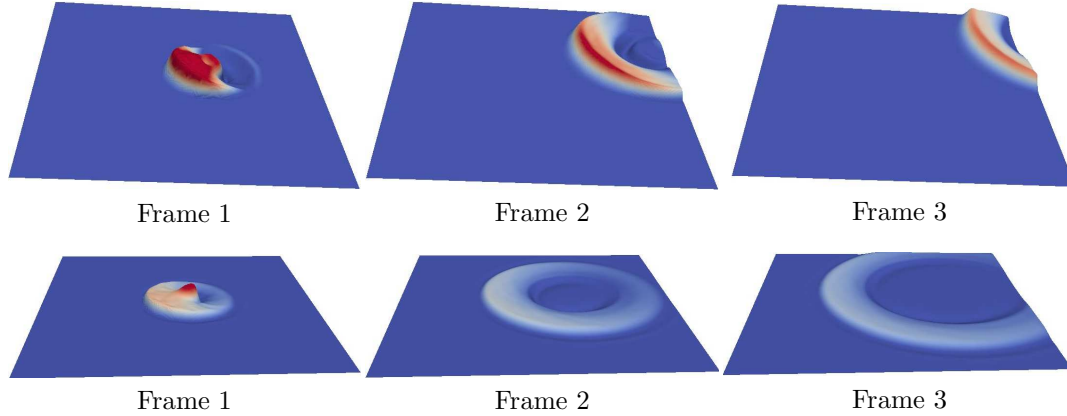


Figure 6.5: Perturbation of channel flows with friction. Super-critical (top) and sub-critical (bottom) flows. Solutions of the LLFs-RK2 scheme (depth)

### 6.3.3 Flows in sloping channels with transverse bed variations

We consider a more general solution obtained when the bathymetry has a constant slope in one direction, and an arbitrary variation in the orthogonal one. The following steady equilibrium is easily found

$$\begin{aligned}
 \eta(x) &= \eta_0 - \xi_0 x & b(x, y) &= b_0 - \xi_0 x + \beta(y) \\
 d(y) &= d_0 - \beta(y) & \text{with } \eta_0 &= b_0 + d_0 \\
 u(y) &= \frac{d(y)^{2/3} \sqrt{\xi_0}}{n} & v &= 0
 \end{aligned} \tag{6.20}$$

Instead of repeating the grid convergence analysis made for the constant energy flows, we show one application in which exact preservation is obtained by *cheating*, and solving the equations on a structured flow aligned grid (left picture on figure 6.6. The standard approximation based on piecewise linear free surface, bathymetry (and hence depth), and discharge allows to *exactly* reproduce the condition  $\partial_y \eta_h = 0$ , thus immediately guaranteeing that no spurious transverse velocities are obtained (see [RAD07] for details). The crosswind coupling being absent, the stream-wise direction is then exactly approximated, as guaranteed by proposition 6.3.4.

Two solutions of (6.20) are computed, one corresponding to super-critical flow, the other to sub-critical flow with. The bathymetry is visible on the left picture on figure 6.6. We perturb the depth in the exact initial state, as shown on the same picture. The time dependent evolutions computed with the LLFs-RK2 scheme are reported on figure 6.7, showing again a nice capturing of the perturbation, and the (exact) preservation of the steady equilibrium away from it.

**Remark 6.3.5** (Cross-wind slope approximation). *As in the case of the constant energy pseudo 1d flows, in this case the important point is the correct approximation of the cross wind slope, which suggests, for future developments, the investigation of ad-hoc cross wind correction terms.*

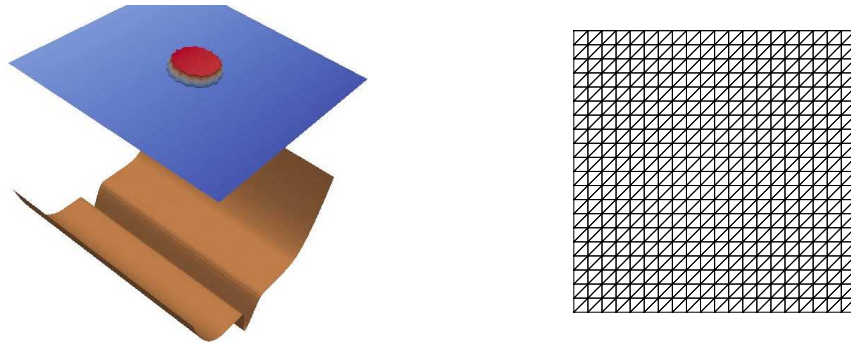


Figure 6.6: Left : Perturbation of 2d sloping channel solution. Left : initial state (free surface). Right : structured triangular grid.

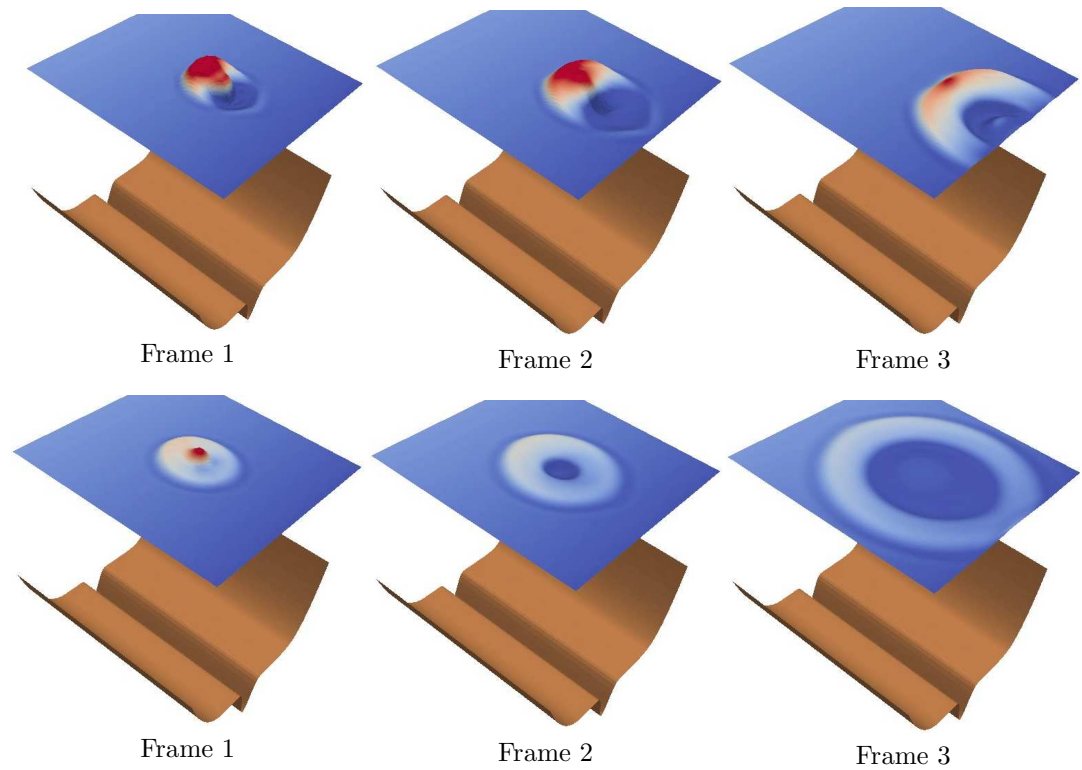


Figure 6.7: Left : Perturbation of 2d sloping channel solution. Super-critical (top) and sub-critical (bottom) case. Solutions of the LLF's-RK2 scheme (free surface)

## Chapter 7

# Contributions : wetting/drying with residual schemes

We consider at last the issue of wetting/drying. We recall the positivity conditions of different type of schemes, and comment on how we implement in practice the schemes such that positivity preservation is indeed achieved. We then recall the modification of the *numerical slope* used to preserve the lake at rest state, and some other computations details such as velocity computation from conserved quantities.

Finally some numerical results and comparison with exact solutions and experiments are shown.

### 7.1 Positivity preservation conditions

As the developments of chapter 4. have shown, obtaining simultaneously positivity and high order of accuracy might require the satisfaction of a time step limitation.

As a result of these developments, we can summarize the positivity preservation properties as follows

**Implicit schemes with CN time integration** The implicit nonlinear LLF-CN scheme (cf. section §4.3 and §4.3.1) can be shown to preserve the positivity of the depth  $d$  provided that [RB09b]

$$\Delta t \leq 2 \min_{K \in \Omega_h} \frac{|K|}{3\alpha}, \quad \alpha > h_K \sup_{\vec{x} \in K} (\vec{v}_h + \sqrt{g d_h}) \quad (7.1)$$

and provided that the limiting is performed equation by equation. A proof is given in [RB09b] to which we refer for details. In the same paper, the scheme based on the BDF2 time integrator introduced in section §4.3.1 is analyzed and shown not to allow any control on the sign of the depth ;

**Explicit RK2-RD scheme** The explicit LLFs-RK2 scheme (cf. section §4.4) can be shown to preserve the positivity of the depth  $d$  provided that

$$\Delta t \leq \min \left( \min_{i \in \Omega_h} \frac{|C_i|}{\sum_{K \in K_i} \alpha}, \min_{K \in \Omega_h} \frac{|K|}{3\alpha} \right), \quad \alpha > h_K \sup_{\vec{x} \in K} (\vec{v}_h + \sqrt{g d_h}) \quad (7.2)$$



and provided that the limiting is performed equation by equation. The proof is similar to the one reported for the CN scheme in [RB09b] and is available in [Ric] ;

**Space time scheme with discontinuous representation in time** The Limited variant of the LF scheme (cf. section §4.2.3)

$$\begin{aligned}\bar{\Phi}_i^{K_t} &= \frac{1}{6} \left( \Phi^{K_t} + \alpha \sum_{j \in K_t} (u_i - u_j) \right) \quad \text{for node } i \text{ at } t^{n+1,-} \\ \bar{\Phi}_i^{K_t} &= \frac{1}{6} \left( \Phi^{K_t} + \alpha \sum_{j \in K_t} (u_i - u_j) \right) \quad \text{for node } i \text{ at } t^{n,+} \\ \Phi_i^{K^-} &= \frac{K}{3} (u_i^{n+} - u_i^{n-}) \quad \text{for node } i \text{ at } t^{n,+}\end{aligned}$$

preserves the positivity of the depth  $d$  *unconditionally w.r.t. the time step* provided that the limiting is performed equation by equation, and that

$$\alpha \geq |K| + h_K \Delta t \sup_{\vec{x} \in K} (\vec{v}_h + \sqrt{g d_h})$$

The proof is easily obtained by showing that the last definition of  $\alpha$ , and the property (2.45) of the limiter allow to put the scheme into the hypotheses of proposition 4.2.3.

This summary shows the interest in the development of the explicit and of the space time schemes as compared to the implicit ones based on Crank Nicholson time stepping.

## 7.2 Wetting/drying and bathymetry approximation

The summary of the previous section tells us that, in order to guarantee the preservation of the positivity of the depth, the limiting (3.4) -(3.5) -(3.6) should be replace by a componentwise procedure.

Unfortunately, limiting using the projection (3.4) turns out to be much more accurate in general [RB09a]. This means that an *ad-hoc* treatment of front cells is needed.

In particular, we summarize hereafter the ensemble of modifications made to the schemes in proximity of dry regions :

**Dry nodes marker and velocity computation** As discussed in detail in [RB09b], a mesh dependent cut off  $C_{h-\bar{v}}$  if sufficient to ensure that the definition

$$\bar{v} = \begin{cases} \frac{\bar{d}}{d} & \text{if } d \geq C_{h-\bar{v}} \\ 0 & \text{otherwise} \end{cases}$$

provides reasonable values for the velocity. In [RB09b] the definition

$$C_{h-\bar{v}} = \frac{h^2}{\max_{i,j \in \Omega_h} \|\vec{x}_i - \vec{x}_j\|}$$

is proposed and extensively tested. Concerning the detection of dry nodes,  $d$  is considered to be *wet* if  $d > \text{eps}_m$  with  $\text{eps}_m$  the machine precision ;

**Limiters switch and stabilization term** A switch is introduced to pass from the projection (3.4) to componentwise limiting. The switch is based on the condition

$$d_m = \min_{j \in K} d_j - \phi_d > 0$$

where  $\phi_d \geq 0$  is an upper bound to the cell mass flux obtained when using the procedure (3.4) - (3.5) - (3.6). In addition, the smoothness sensor  $\delta(u_h)$  (cf. item 3.a section §4.3.1) in the stabilization is smoothly turned off when  $d_m$  approaches zero. When using total energy variables, the same switch is used to change the interpolation to conservative variables (which simply implies using directly  $d$  instead of the total energy). The interested reader can refer to [RB09b] for more details ;

**Numerical slope** Lastly, the numerical slope is redefined following the initial proposition of Brufau and Garcia Navarro [45] :

$$\nabla b_h^* = \sum_{j \in K} b_j^* \varphi_j$$

with  $b_j^* = b_j$  if  $d_j > \text{eps}_m$ , otherwise

$$b_j^* = \max_{\substack{l \in K \\ d_l > \text{eps}_m}} \eta_l$$

This modification is easily shown to guarantee the preservation of the lake at rest in presence of dry areas. When using analytical bathymetries, the switch on  $d_m$  is used to locally revert this approximation.

## 7.3 Results

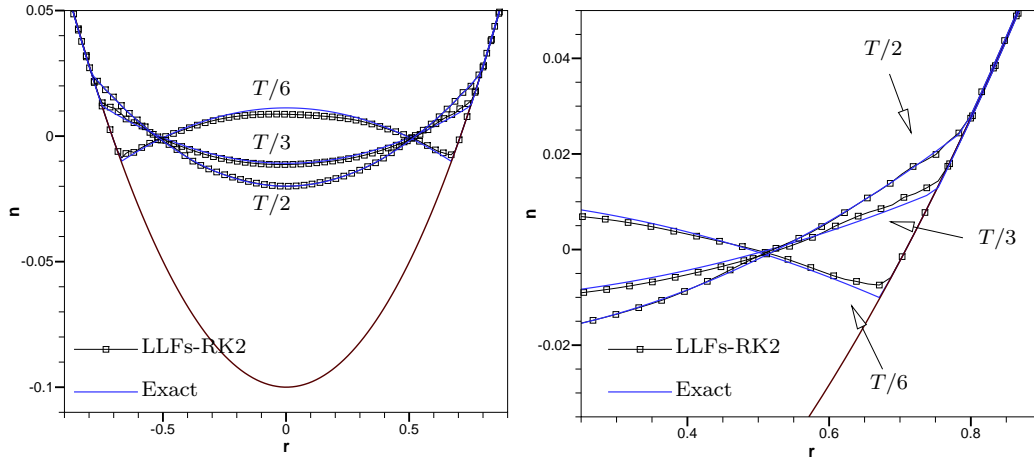


Figure 7.1: Thacker's curved oscillations. Data along line the  $y = 0$ . LLFs-RK2 scheme

### 7.3.1 Thacker's oscillations

We present results on three tests involving dry areas. The first is the well known analytical solution of Thacker involving oscillations in a paraboloid basin. We refer to [248, RB09b] for the description of the test. Here we consider the case of a curved free surface and we solve the problem on unstructured grids with the topology shown in the rightmost picture on figure 6.3. We have run computations with the LLFs-CN and with the LLFs-RK2, and performed a grid convergence study. For these tests, the approximation is written in conservative variables.

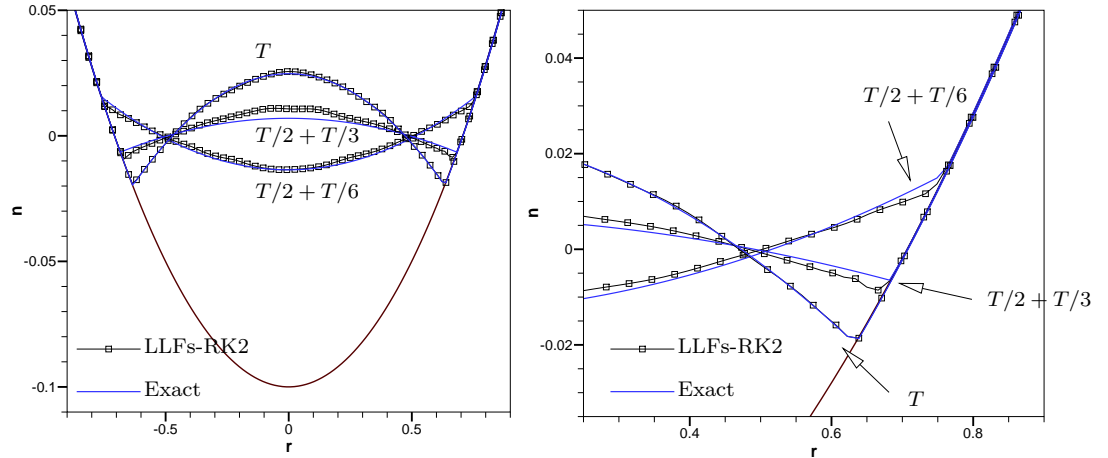


Figure 7.2: Thacker's curved oscillations. Data along line the  $y = 0$ . LLFs-RK2 scheme

The results of the two schemes are qualitatively very similar, and on figures 7.1 and 7.2 we report the free surface line plots along the  $x$  axis for the LLFs-RK2 solutions on the mesh obtained after one conformal refinement ( $h \approx 0.05$ ). The numerical profiles nicely approach the exact ones without any under/overshoot, showing the effectiveness of the wetting/drying procedure used.

**CPU time for one period**  
 (on a MacBook Pro laptop  
 with 2.66 GHz processor)  
 LLFs-RK2 : 1m43.470s  
 LLFs-CN : 13m33.110s

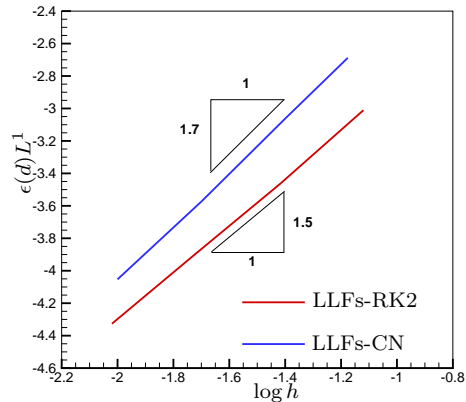


Figure 7.3: Thacker's oscillations : grid convergence. LLFs-RK2 and LLFs-CN schemes

On figure (7.3) we report the plot of the convergence of the  $L^1$  error on the free surface after one period of oscillation. The errors of the explicit LLFs-RK2 scheme are lower than those of the implicit scheme which, however, gives a better convergence rate. Next to the convergence plot we have reported the computational time required for one full period, showing that the explicit scheme is about ten times faster than the scheme based on implicit CN time stepping. This is a direct consequence of the fact that the two schemes verify time step limitations very similar even if the second is implicit (cf. equations (7.1) and (7.2)).

### 7.3.2 Wave run up on a conical island

The second set of results reproduces the experiments of [44] involving the run up of a long wave on a conical island. Detailed description of the test case can be found in [RB09b]. We report here the results obtained for one of the experimental conditions, characterized by a wave amplitude of 20% w.r.t the undisturbed level. The Manning coefficient is set to  $n = 0.014$ . We report the results obtained with the LLFs-RK2 scheme. We have used the approximation written in total energy variables with a piecewise linear approximation of the bathymetry, to show feasibility of the use of this approximation in flows containing dry areas.

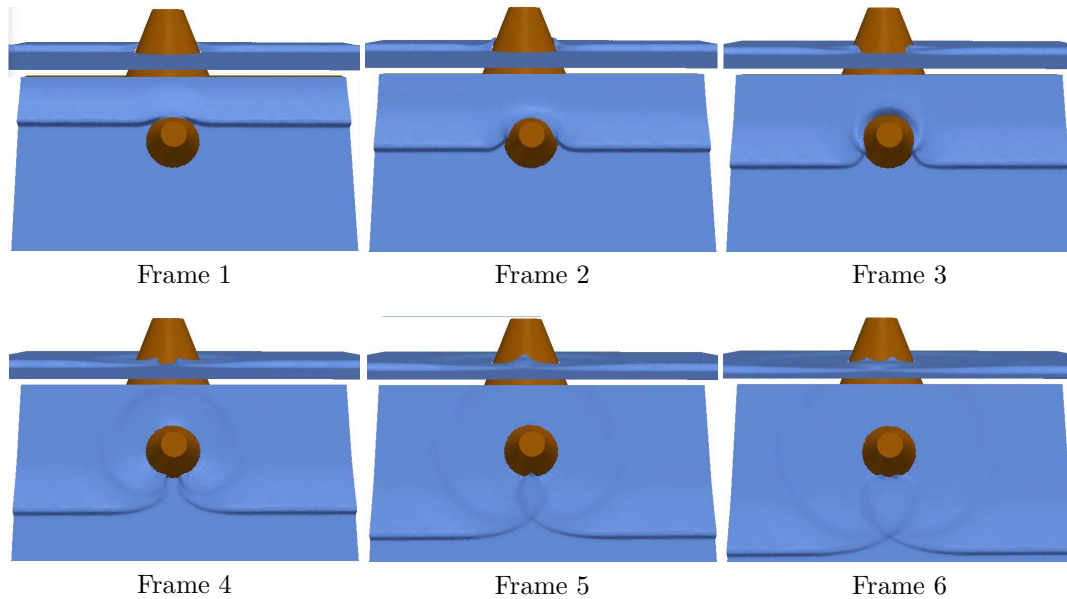


Figure 7.4: Wave runup on a conical island. Flow overview. Solution of the LLFs-RK2 scheme. In every frame : on top the rear view and on the bottom a top view

On figure 7.4 we report a 3d visualization of the flow showing the nice capturing of the interaction and of the run up on the rear side, clearly visible in Frame 5. We then consider the run up plot obtained by constructing the locus of highest run up points throughout the flow, and compare it with the experimental one. The comparison is reported on figure 7.5 showing an excellent agreement. Lastly, we consider the time evolution of the free surface in the gauge locations indicated on the right picture on the same figure. The comparison of the computed time history with the data of [44] shows an excellent agreement. The interested reader can consult [RB09b] for similar results obtained with the LLFs-CN scheme.

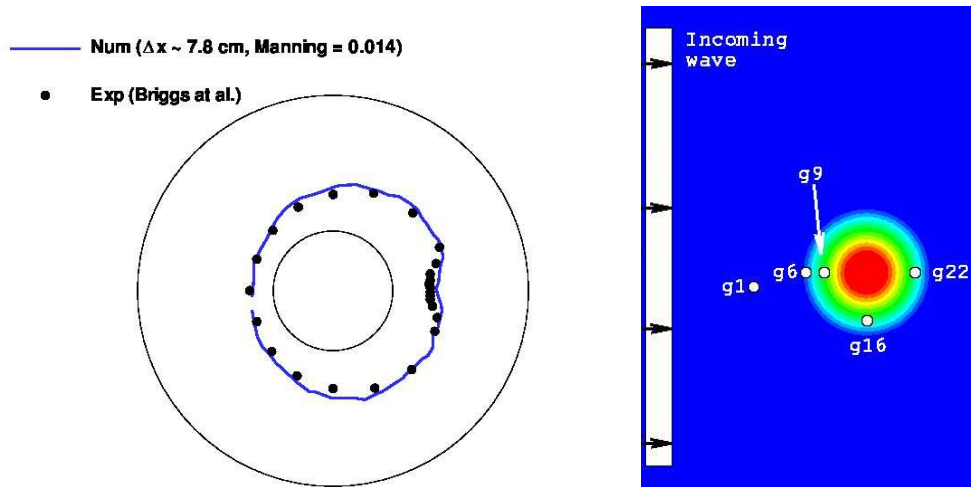


Figure 7.5: Wave runup on a conical island, LLFs-RK2 scheme. Left : runup plot, comparison with the experiments of [44]. Right : gauge locations

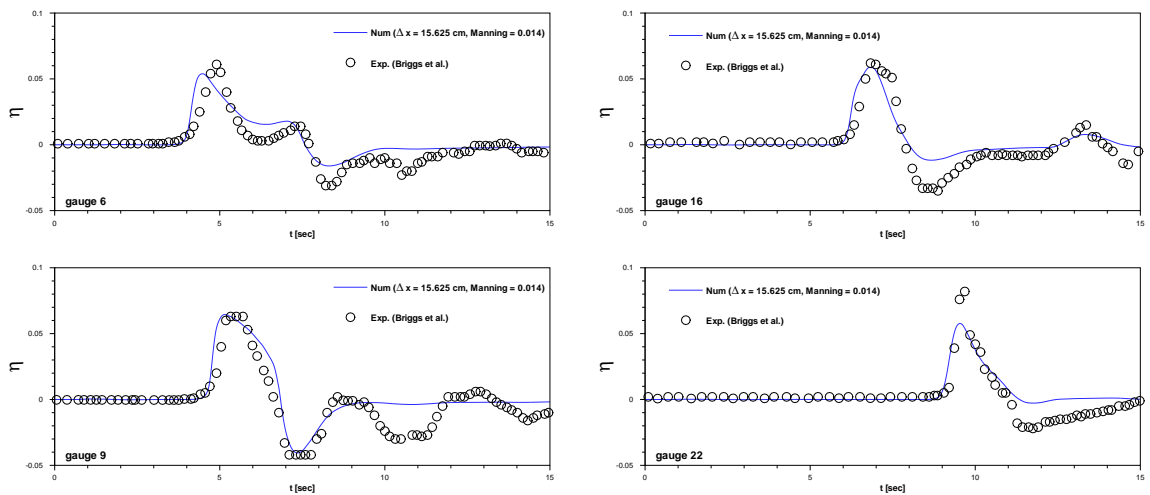


Figure 7.6: Wave runup on a conical island, LLFs-RK2 scheme. Comparison of free surface position with experiments of [44]

### 7.3.3 The 1993 Okushiri tsunami test case

This test involves the interaction of a long wave with a complex three dimensional bathymetry. It is one of the tests proposed in the third international workshop on long-wave run up models. A detailed description, and the data describing the bathymetry, the incoming wave height, and data sets of experimental measures taken on a laboratory model of the flow are available on the web site :

[http://isec.nacse.org/workshop/2004\\_cornell/bmark2.html](http://isec.nacse.org/workshop/2004_cornell/bmark2.html)

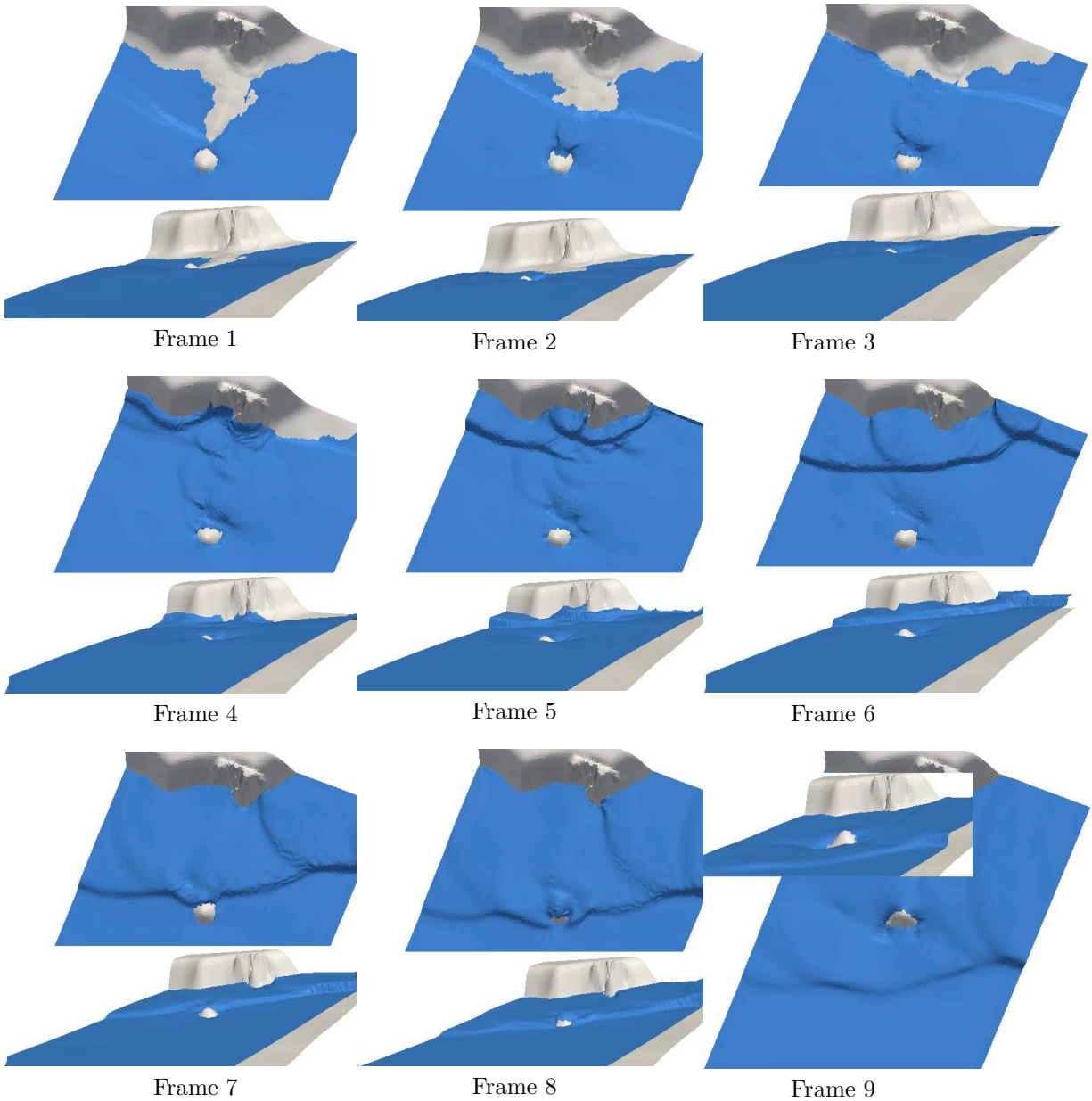


Figure 7.7: Okushiri tsunami simulation. Flow overview. Solution of the LLFs-RK2 scheme. In every frame : on top a view from the top and on the bottom a side view

We report on figure 7.7 a 3d visualization of the results obtained with the LLFs-RK2 scheme. The pictures show the incoming wave reaching the shore and the complex system of reflections arising from the interaction, including the run up on the island. Finally, figure 7.8 reports the comparison of the time evolution of the free surface in the gauge locations indicated on the picture on the top left in the same figure.

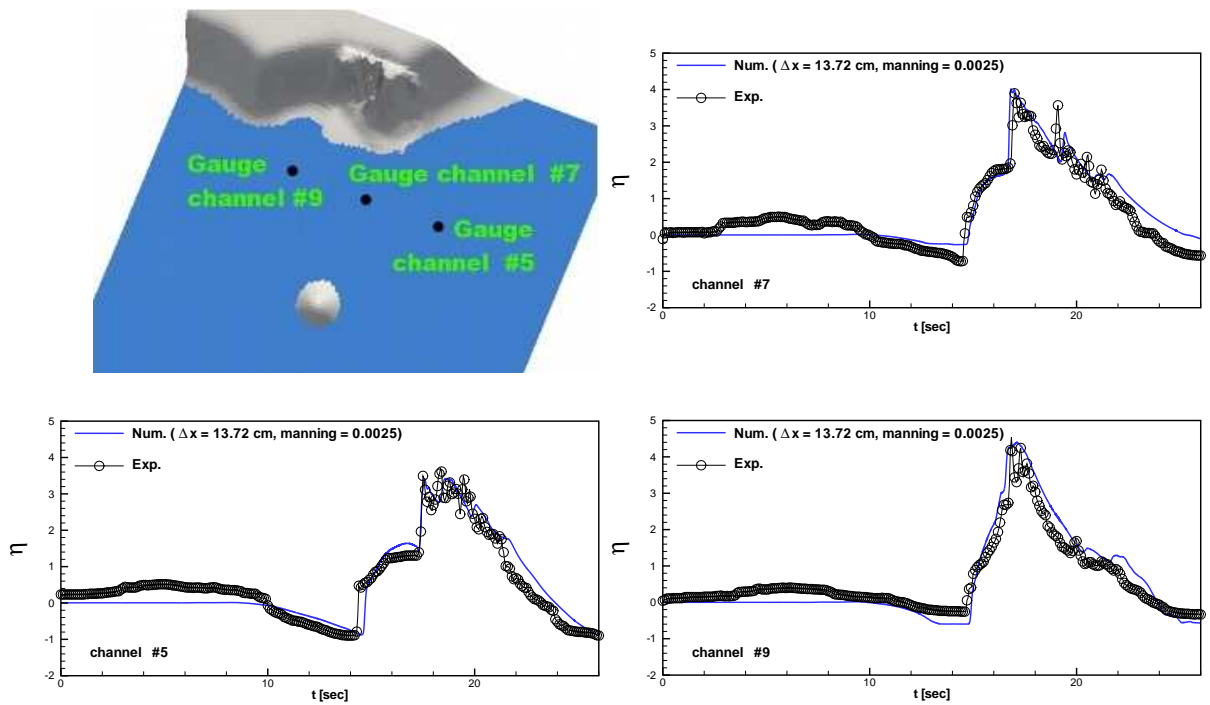


Figure 7.8: Okushiri tsunami simulation. Gauge positions and comparisons with data of the workshop

## Part III

# Conclusions and perspectives





## Chapter 8

# Summary of contributions and perspectives

After a very short summary of the scientific contributions discussed in this manuscript, in this chapter the current and planned (mid-term) research activities of the author are discussed, as they are foreseen today.

Needless to say, this discussion only represents a rough roadmap of an activity that can (and certainly will) change accordingly to the results obtained in the course of the years.

### 8.1 Summary of contributions

This manuscript summarizes my work on the construction of residual based discretizations for the approximation of solutions to hyperbolic conservation laws, and their application to shallow water simulation.

The presentation has (hopefully) put this research into context, and shown the motivation behind the work, and the main objectives it aims at. The main achievements described in the manuscript can be listed as follows :

- Development of a general conservative framework for residual schemes based on a continuous direct approximation of the flux. This framework allows the construction of arbitrary order schemes based on the interpolation of any set of variables. This allows to free the discretization of the need of a Roe parameter, at the same time allowing to choose the set of approximation variables based on physical arguments ;
- Development of higher order variants of the so-called residual distribution schemes employing higher degree Lagrange polynomial approximation in space. This part of the research has radically changed the way in which residual distribution schemes are constructed. The weakness of the classical compute balance-distribute have been underlined, and a solution proposed, giving new directions for future research ;
- Study and development of residual schemes for time dependent flows. A thorough characterization of consistency and accuracy has allowed to give the tools to construct arbitrary order schemes for time dependent problems. In particular, whole families of implicit schemes multistep time integration, and of genuinely space time schemes have been studied, implemented and tested ;

- Development of a framework for constructing genuinely explicit and simplified variants of truly residual methods. This approach allows to greatly simplify the mass matrix, reducing it to that of the Galerkin scheme, or of a centered finite volume one. Indeed, while particularly interesting for the residual distribution schemes developed by the author, the result has application also to Galerkin finite elements schemes with bubble stabilization (*e.g.* SUPG or GLS) and to residual based finite volume discretizations ;
- Investigation of the numerical preservation of multidimensional steady equilibria in shallow water flows. The concept of super-consistency allows to generalize the concept of C-property to approximations on general meshes when a sufficiently regular description of the bathymetry is available. The idea is tested and validated on several steady equilibria, including some not known in literature so far ;
- Proposition of accurate residual-based positivity preserving schemes for the simulation of complex free surface flows on unstructured meshes, This gives an alternative approach to the construction of adaptive discretizations for free surface flows on complex bathymetries .

These results have brought the class of schemes known as residual distribution to a much higher level of maturity. At least in the hope of the author, these developments may also have given some more general contributions that can be applied in other contexts.

On the other hand, the work discussed here leaves some questions open, leading to some new open perspectives. These are the basis for the author's current and planned research activities, and are discussed in the following sections.

## 8.2 Perspectives : residual schemes for conservation laws

The work of the last years has led to some results that change quite radically the way in which the so-called residual distribution schemes are conceived. These are quite general issues which involve both the steady and time dependent case, as well as any adaptation of the schemes to a particular class of problems. Part of the current (and foreseen) research of the author deals with these issues and is discussed hereafter.

**Petrov-Galerkin schemes based on RD techniques.** As proposition 3.2.4 shows, the paradigm “compute balance-distribute” works only on linear triangles. In every other case some sort of sub-elemental resolution is necessary to avoid the emergence of spurious modes. For this reason, some of the current research is devoted to the construction of schemes based on the Petrov-Galerkin (PG) form of  $\nabla \cdot \mathcal{F} = 0$  :

$$\sum_{K \in \Omega_h} \int_K \beta_i^K(u_h, x, y) \nabla \cdot \mathcal{F}_h(u_h) = 0 \quad \forall i \in \Omega_h$$

Two questions arise. The first is how to define the PG test functions  $\beta_i^K(u_h, x, y)$  such that the properties of the RD schemes constructed so far are recovered. The second is : given the definition of  $\beta_i^K(u_h, x, y)$ , which can be complex and present a strong nonlinear dependence on the discrete solution, how can we smartly guarantee discrete conservation without having to compute derivatives of  $\beta_i^K(u_h, x, y)$ .

This questions have been already (at least partly) answered as part of the work in progress in collaboration within the INRIA team BACCHUS (with R. Abgrall), with the group led

by H. Deconinck at the von Karman Institute (co-supervision of the PhDs of M. Vymazal and S. D'Angelo), and with A. Larat at the École Central Paris [VRQD11, DRAD11, LR]. The answers are based on two simple ideas :

- for steady state scalar constant advection, for example, instead of considering an integrated residual one can consider the local residual  $\vec{a} \cdot \nabla u_h$ , and decompose this quantity as done in section §2.2.4 for the fluctuation :

$$r_h^K = \sum_{j \in K} \vec{a} \cdot \nabla \varphi_j(x, y) u_j = \sum_{j \in K} k_j(x, y) u_j, \quad k_j(x, y) = \vec{a} \cdot \nabla \varphi_j(x, y)$$

At this point any known distribution strategy can be applied. In particular, test functions  $\beta_i^K(x, y)$  are obtained by using known accuracy preserving schemes.

- when dealing with nonlinear conservation laws, discrete conservation must be ensured. This issue is dealt with an approach similar to the one used in the context of the spectral finite difference method [266, 177] : we construct higher order continuous local polynomial approximation of the flux  $\mathcal{F}_h$ , so that we can replace

$$\nabla \cdot \mathcal{F}(u_h) \approx \nabla \cdot \mathcal{F}_h(x, y) = \sum_{\sigma \in K} \mathcal{F}(u_h(x_\sigma)) \cdot \nabla \bar{\varphi}_\sigma(x, y)$$

where in general the points  $x_\sigma$  (and the corresponding basis functions  $\bar{\varphi}_\sigma$ ) define a polynomial approximation of degree higher than that of  $u_h$ . The continuous PG statement is finally replaced by the fully discrete one

$$\sum_{K \in \Omega_h} |K| \sum_{q=1}^{Q_p} \omega_q \beta_i^K(u_h(x_q, y_q), x_q, y_q) \nabla \cdot \mathcal{F}_h(x_q, y_q) = 0 \quad \forall i \in \Omega_h$$

and discrete conservation is ultimately guaranteed by the constraints on the test functions and on the quadrature formula :

$$\sum_{j \in K} \beta_i^K(u_h(x_q, y_q), x_q, y_q) = 1$$

$$|K| \sum_{q=1}^{Q_p} \omega_q \nabla \cdot \mathcal{F}_h(x_q, y_q) \stackrel{\text{exactly for } \mathcal{F}_h}{=} \int_K \nabla \cdot \mathcal{F}_h(x, y) = \oint_{\partial K} \mathcal{F}_h(x, y) \cdot \vec{n}$$

**Viscous terms.** The treatment of viscous terms, which has been left out of this manuscript, currently constitutes an important research topic (co-supervision with R. Abgrall of the PhD of G. Baurin and of D. De Santis at INRIA). The main problem is to keep a proper balance between inviscid and viscous discrete operators, such that the accuracy obtained in practice is uniform w.r.t the data or the problem, in particular with the *Peclet* (or *Reynolds*) number. In continuous stabilized Galerkin schemes (*e.g.* SUPG scheme) this is taken care of by weighting the stabilization of the inviscid operator by some Peclet-dependent parameter, chosen on the basis of a rigorous error analysis [110]. A similar approach has been attempted in the residual distribution context [RVAD05, RVAD08], however the lack of sufficient theoretical tools (in particular stability estimates) for a rigorous error analysis, makes this approach difficult to follow. A more interesting method is the one proposed by H. Nishikawa [190, 191], bearing

some similarities with Least-Squares discretizations (see *e.g.* [108]). The basic idea is to manipulate the mixed form of the advection diffusion so that a first order hyperbolic system is obtained for the unknown and its derivatives (viscous fluxes). The system is then discretized with any known residual distribution scheme, obtaining uniformly (w.r.t. the Peclet number) high order results. A more interesting variant of the approach [192] is to use the first order system form of the equations to derive a scheme and to discard the equations for the gradients (viscous fluxes), which are replaced by high order reconstructed values. This is the approach currently investigated within the PhD theses of G. Baurin and D. De Santis [ABSR]. Other variants of the method involve the use of viscous numerical fluxes [62]. These options will be investigated in the future.

**(Locally) Discontinuous approximation.** In some cases it is convenient to be able to handle correctly locally edge-discontinuous spatial approximations. This happens, for example, in all applications requiring non-conformal meshes, or to be able to effectively use both  $h$ - and  $p$ - adaptation techniques. Some initial constructions allowing to handle discontinuous approximation with RD schemes have been discussed in [61, 140, 12, 3]. For steady conservation laws, the simplest way to describe most of these developments is, similarly to what is done in the space-time case in section §4.2.3, to introduce additional fictitious edge cells and to add to the volume terms, the face fluctuations  $\phi_i^f$  such that

$$\sum_{j \in K} \phi_j^f = \int_f [\mathcal{F}_h(u_h)] \cdot \vec{n}_f$$

[.] denoting the jump. The schemes thus obtained bear some similarities with Discontinuous Galerkin. The investigation of these relations, and the further development of the work of [140] are the object of the PhD of A. Warzynski (University of Leeds) that the author is co-supervising in collaboration with M Hubbard.

**Error estimation and adaptation.** An important ingredient to fully profit from higher order discretization methods is the availability of an error estimation technique allowing to efficiently adapt the mesh and/or the degree of the polynomial approximation. A powerful tool giving such estimators is the theory of the discrete adjoint, which has been quite successfully used in the context of DG schemes [127, 128, 129]. The development of similar techniques in the framework of Petrov-Galerkin discretizations, including the *RD flavored* Petrov-Galerkin schemes developed in [VRQD11, DRAD11, LR], is the objective of the PhD thesis of S. D'Angelo that the author co-supervises with H. Deconinck (von Karman Institute for Fluid Dynamics). Independently on the details, for which we refer to the preprint [DRAD11], the basic idea is that for a continuous approximation in space one can associate to the direct problem (steady scalar advection)

$$\text{find } u_h \text{ such that } \sum_{K \in \Omega_h} \int_K \beta_i(x, y) \vec{a} \cdot \nabla u_h = r_i^{\text{B.C.s}} \quad \text{with} \quad u_h|_K = \sum_{j \in K} \varphi_j(x, y) u_j$$

the adjoint problem

$$\text{find } \tilde{z}_h \text{ such that } - \sum_{K \in \Omega_h} \int_K \tilde{z}_h \vec{a} \cdot \nabla \varphi_i(x, y) = r_i^{\text{Adj}} \quad \text{with} \quad \tilde{z}_h|_K = \sum_{j \in K} \beta_j(x, y) z_j$$

Due to the limited regularity of the  $\beta_i(x, y)$  functions, this adjoint problem is generally well posed only in  $L^2$ , and the discrete adjoint solution  $\tilde{z}_h$  turns out to be only a first order approximation of the exact one. Moreover, the approximation is clearly not self adjoint in general (different approximation spaces, the adjoint one having very low regularity), as confirmed by the absence of super-convergence on functionals. Nevertheless, this framework allows the formal construction of a local approximation of the adjoint solution, and thus of a local goal oriented estimation of the error that can be used for adaptation purposes. Examples and generalizations are discussed in the preprint [DRAD11]. The objective of the PhD of S. D'Angelo is to further develop the framework toward compressible aerodynamics applications.

**Finite element spaces.** So far most implementations of the residual method discussed in the manuscript make use of standard  $P^k$  and  $Q^k$  Lagrange elements, the only exception being [13]. Several other choices will be extensively tested and compared. Examples are the Bezier polynomials used in [13], as well as enriched Lagrange elements allowing for mass lumping [72, 156], and continuous hierarchical elements (see *e.g.* [270] and references therein) which might simplify the implementation of  $p$ - adaptive versions of the method.

**Scalar decompositions and preconditioning.** For systems of conservation laws, all the work discussed in this manuscript is based on a compact matrix formulation. One of the author's foreseen developments is to replace this formalism by a set of similarity transformations, allowing to at least approximately decouple (exact decoupling being obtained only at the differential level) some scalar components of the system (typically at least the entropy equation). This will greatly reduce the computational cost of the schemes. In particular, the work made in the past on local preconditioning and approximate diagonalization [202, 36, 210] will be revived and extended in order to : reduce the cost of higher order schemes ; improve the accuracy in low Mach regions for higher order schemes ; simplify and reduce the cost of the discretization in the unsteady case.

**Stabilization operators.** The nonlinear schemes currently implemented in the code of INRIA make use of the construction discussed in section §3.2.3 based on the paradigm : compute low order positive local update - apply limiter (to obtain formally high order positive local update) - add dissipative operator in smooth regions (to recognize/kill spurious modes and achieve optimal iterative/grid convergence). So far, the last step involves the use of a least squares type bi-linear operator similar to that used in SUPG/GLS discretizations [108, 21]. This term is associated to each element, over which is evaluated as a linear combination of pointwise local operators. A different approach could be to exploit the  $C^0$  nature of the approximation to introduce face based corrections that depend on the jumps of the first order derivatives of the discrete solution, as proposed in [50, 49]. This approach might turn out to have some advantage in terms of locality. In this context, a long term research direction will be the further understanding of the algebraic properties associated to the Multidimensional Upwinding procedure (cf. section §2.2.4). The objective is, ideally, to be able to reproduce the parameter free, stable and accurate behavior of the  $P^1$  schemes in more general contexts.

### 8.2.1 Schemes for unsteady conservation laws

The development of more efficient and more accurate residual based schemes for time dependent conservation laws remains at the core of the activity of the author. Most of the topics mentioned in the previous section have a direct impact on this activity. Some other topics

have particular interest in the time dependent case. The main issues under investigation and some foreseen developments are discussed hereafter.

**Higher Order in space and time.** The extension to higher orders of both the explicit Runge-Kutta based schemes, and of the space-time schemes with discontinuous time approximation is currently being studied [AR, LR] in collaboration with R. Abgrall at INRIA and with A. Larat at École Centrale Paris. As discussed in sections §4.2, §4.4, and §7.1, the objective is to provide higher order residual based constructions allowing the preservation of the non-oscillatory character of the solution, and of its positivity, either unconditionally on the time step, in the case of the implicit space time schemes, or of a genuinely explicit nature, if the positivity preservation induces a time step limitation.

**Choice of the finite element space.** A critical point in the development of schemes for type dependent conservation laws is the choice of the approximation ins space. This choice influences both the meaning of the positivity preservation property, and the mass lumping procedure. An example of approximation spaces that show similar mass lumping properties is given by the Bezier polynomials used in [13] and the enriched Lagrange elements developed *e.g.* in [156]. In the two cases, however, the degrees of freedom do not represent the same quantity : in the Bezier case they are solution values in control points, in the Lagrange case they are interpolated values of the solution. This radically changes the meaning of the positivity property, and its implications on the behavior of the polynomial approximation. These issues are under study in collaboration with R. Abgrall at INRIA [AR]. In addition, the use of discontinuous approximations in the context of explicit residual based discretizations is being studied in the PhD of A. Warzynski (University of Leeds) that the author is co-supervising in collaboration with M Hubbard.

**Solution techniques for space time schemes.** An extremely important topic is the improvement of the solution strategies used to solve the nonlinear equations in the context of nonlinear space-time discretizations. As recalled more than once, the advantage of these schemes is the possibility of having an unconditionally higher order accurate and positive scheme. The term *unconditionally* has to be intended as *for arbitrarily large time steps*. The possibility of increasing the time step (*viz.* CFL number) size is what should overcompensate the additional cost of solving a space-time problem. This is true only if the solution at the new time step can be obtained in an efficient manner. Explicit pseudo-time stepping loops currently implemented [DRD03a, DRD05, HR11] can require a large number of iterations to converge thus becoming very ineffective, if the iteration parameter is not properly *tuned*. On the other hand, implicit methods, while converging in less iterations, become quickly time-consuming (and memory consuming in the higher order case) due to the necessity of assembling an approximate Jacobian of the equations [DRD03a, DRD05]. These techniques should be replaced by some smarter strategy allowing a finite (and as small as possible) number of cheap pseudo-time iterations. Techniques that will be investigated are either based on ad-hoc Runge-Kutta integrators, as in [158], or on efficiently designed predictor-multicorrector methods, as in [161, 162]. Another possibility that will be investigated is the use of very few Newton-Krylov iterations in conjunction with matrix-free formulations (see *e.g.* [159, 28] and references therein).

**Fourier analysis (linear schemes).** A key point, which has not been addressed so far, is the accuracy of the schemes in terms of dispersion. This point is critical, the correct re-

production of the physics of wave propagation requiring a very small dispersion error. Some initial studies in this direction have been made at the von Karman Institute for fluid Dynamics [160] for the space-time LDA scheme and the LDA scheme plus mass matrix discussed in sections §4.2.2 and §4.3 (see also section §4.4.1). A similar study will have to be undertaken both for the higher order variant of the space-time schemes with discontinuous representation in time [HR11, LR], and for the explicit schemes [RA10, AR].

**Explicit time integration techniques.** The work made so far is based either on time integrators of the Runge-Kutta type, or on more classical implicit multi-step schemes, or on fully space-time approaches. A path that has never been followed is a combination of time integration of the PDE combined with a truncated Taylor series development in time. This could allow to increase the temporal accuracy, and to derive in one shot the spatial and temporal discretization, without having to manufacture or look for stable higher order time integration formulae. The key point is how to include the time derivatives obtained in the Taylor expansion phase. This approach is very similar to what is done in the context of continuous finite elements in Taylor-Galerkin discretizations [99, 100]. In the framework of finite volume and discontinuous Galerkin schemes, very interesting developments have been reported during the last ten years in [226, 103, 104, 114, 115, 172]. The challenge is how to use similar ideas in the context of a global continuous approximation, and with a nonlinear discretization strategy.

### 8.3 Perspectives : free surface flows and other activities

Aside from schemes development, the author of the manuscript is involved in a number of activities concerning the simulation of free surface flows, and the numerical modeling of oxidation and healing phenomena in composite materials.

This work is concerned on one hand with the improvement and adaptation of residual based schemes, in order to correctly reproduce the physics described by a given model, and on the other with the study and eventually development of the model themselves. A discussion of these activities is given in the following sections.

#### 8.3.1 High order numerical modeling of free surface flows

The simulation of free surface flows constitutes, in the mind of the author, one of the most promising future perspectives and applications of the schemes described in the manuscript.

The transfer of numerics born in the aeronautics and aerospace community to civil, hydraulic, coastal and environmental engineering is an important direction in which the author will invest himself in the following years. Applications going from hazard forecast/simulation (flooding related to violent rainfall or to tidal waves, tsunami-coastal structure interactions, efficiency of wave-breaking structures etc.), to transport of pollutants in complex hydrological networks, to the efficiency of irrigation systems, might profit from the use of advanced adaptive techniques developed in a more industrial context.

Concerning instead the ongoing and foreseen short term activities, the most relevant issues are discussed hereafter.

**Higher order schemes, stiffness and time integration.** The natural future development of the work discussed is the use of higher order space-time or genuinely explicit schemes to



solve the Shallow Water equations. This work has partially started in collaboration with the University of Leeds [HRS11, SHR] for the space-time schemes. The most interesting issue that still remains to be assessed is the efficiency one : which approach leads to the least time-consuming strategy?. The answer will of course depend on the developments of higher order schemes for time dependent flows discussed in the previous section.

**Mesh and polynomial adaptation.** Adaptation techniques are a necessary ingredient to provide an efficient and accurate simulation tool. For free surface flows, mesh adaptation techniques based on the regularity of the bathymetry, combined with some time-dependent adaptation strategy taking into account the position of the wet/dry front and other criteria seem the most reasonable way to proceed. The unstructured mesh approach discussed in the manuscript gives a perfect framework for this application. Work in this direction is planned with C. Dobrzynski of the INRIA team BACCHUS.

**Viscous and dispersive terms.** For many applications, especially in coastal engineering and in the simulation of tidal flows in rivers, the basic Shallow Water does not provide a description accurate enough. Higher order terms, modeling diffusive and dispersive components of the flow, need to be taken into account. A large literature exists on free surface models including these corrections (see [37, 165] for a discussion). A collaboration with P. Bonneton of the EPOC<sup>1</sup> research unit in Bordeaux, F. Marche of the University of Montpellier, and with H. Beaugendre of the INRIA team BACCHUS is foreseen to develop ad-hoc unstructured mesh techniques to deal with the nonlinear Serre-Green-Naghdi (SGN) models proposed in [37, 165]. Further collaborations is planned with the EPOC research unit on the transition from the full SGN model to the Shallow Water one in near shore simulations. In addition to these collaborations, the author is currently investigating the relations between the super-consistent residual based approach discussed in the manuscript and finite volume schemes for a one-dimensional viscous shallow water model with friction. The objective of this study is to use the residual approach to construct finite volume fluxes verifying the super-consistency property.

**Friction laws and model reliability issues.** Free surface models based on vertical averaging and asymptotic expansions, as those in use in the hydraulic and coastal engineering community, very often contain parameters modeling multidimensional effects which are *lost* in the averaging/asymptotic expansion process. The physics described by these models is often dependent on these parameters, and eventually on the form of sub-models in which these parameters figure. A typical example is the friction terms in Shallow Water flows. This term, which has a tremendous impact on the wetting/drying process, hence on flooding and run-up phenomena, admits a number of formulations all depending on parameters more or less available depending on the problem under study. A recent discussion can be found in [185]. In the reference, the accuracy of several models has been studied by comparing to a model experiment. The question is however, besides accuracy, what model is the most reliable, *i.e.* the least dependent on the values of the parameters. To answer this question, a study based on the quantification of the uncertainty of the simulation output w.r.t. the uncertainty on the friction term parameters (and other input data) has been started in collaboration with P. Congedo, G. Geraci and R. Abgrall at INRIA [RCGA12]. The plan is to perform a methodical investigation of the reliability of representative friction models in

---

<sup>1</sup>Environnements et Paléoenvironnements Océaniques et Continentaux

applications involving flooding and run-up. Similar studies are foreseen for more complex models involving the high order terms discussed above.

**Upwind schemes and streamline-crosswind decomposition.** As already discussed in section §8.2, the understanding of the implications of Multidimensional Upwinding on the algebraic properties of the discrete equations might lead to improved schemes. In particular, upwind positivity preserving discretizations for the Shallow Water model might represent an important improvement w.r.t the one based on the Lax-Friedrich's type scheme implemented so far. Additionally, the developments of section §6.3 show that the Shallow Water system admits several steady equilibria which are essentially one-dimensional in the streamline direction. As a consequence, an approach based on a streamline/crosswind decomposition of the system could guarantee the approximate preservation of these equilibria without having to resort to the use of exact/analytical representations of the bathymetry, and to very expensive quadrature formulae;

**Super consistency and real multidimensional flows.** Real multidimensional steady state equilibria for the Shallow Water equations can be easily built (see [RAD07] for an example). Unfortunately, the preservation of these equilibria requires the exact discrete approximation of differential constraints of the solenoidal or irrotational type. The preservation of these constraints at the discrete level is much more difficult than the preservation of some set of invariants. This is however an exciting subject that bears similarities with vorticity and divergence preserving discretizations. The residual-based framework seems to provide a good framework to study such discretizations [169]. Other approaches exist in the nonlinear case, see [109] for an example involving the Shallow Water equations. In the author's mind, however, a very promising idea is the potential based approach suggested in [183, 182]. A residual based view of the schemes proposed in the last reference is one of the topics the author will pursue in the future.

### 8.3.2 Numerical modeling of oxidation and healing processes in composite materials

In addition to the topics discussed so far, the author is engaged since a few years in a modeling activity in the field of advanced composite materials. The objective of this activity is the development of a numerical model for a class of Ceramic Matrix Composites (CMCs), in particular the carbon-based composites referred to as *self healing* composites [120, 64]. In these composites, the matrix surrounding the silicon-carbide fibers has a multilayer structure including several layers of a boron-carbide. At temperatures between 500°C-800°C, typical of civil aero-engines, this boron-carbide oxidizes very fast producing a very viscous liquid polymer. In oxygen-filled cracks, the boron-carbide oxidizes, and the liquid boron-oxide fills the crack. This eventually creates a barrier protecting the composite fibers from the external oxidizing environment, ultimately increasing the lifetime of the material. This is what one refers to as the healing process [187, 206, 207]. The mechanical properties of the material after healing remain sufficiently close to the original ones to make these composites ideal for use in aero-engines, as shown by their use in the technology at the basis of the LEAP-X project of the CFM joint venture [14, 15] <sup>2</sup>.

The lifespan of components built of these composites is such that mechanical tests in the material/component design phase are impractical in most cases. This makes the develop-

<sup>2</sup>see also <http://www.cfm56.com/media/pdf/LEAP.pdf>

ment of numerical models a necessity [64], and has led to the development of a concerted action involving material experts, structural mechanics experts, and applied mathematicians aiming at providing the necessary tools to build a virtual material laboratory for ceramic matrix composites [CVD<sup>+</sup>11]. As part of this effort, the author has established a strong collaboration with the LCTS laboratory<sup>3</sup> in Bordeaux with the objective of providing a numerical model of the physicochemical behavior of CMCs [CVD<sup>+</sup>11, DPC<sup>+</sup>12] (and see also [DRV10a, DRV10b]). The main issues at the core of this collaboration are the following.

**Vertically averaged models.** The main idea of the numerical model under construction is that the mesoscale physicochemical behavior of the material can be approximated as two-dimensional, *i.e.* an average description on the crack thickness is enough. A large initial part of the work consists in deriving crack averaged models for oxygen diffusion and reaction, and for the evolution of the liquid polymer. Similar models are encountered in some two-layer Shallow Water approximations, see *e.g.* [157, 43, 186]. The work done so far [CVD<sup>+</sup>11, DPC<sup>+</sup>12] assumes that the growth rate of the liquid is sufficiently large to neglect the liquid flow in the early phases of the healing process. Average coupled models for liquid flow/oxygen diffusion reaction are being developed in collaboration with the LCTS, in the framework of the co-supervision of the PhD thesis of G. Perrot (Université de Bordeaux I), and with F. Marche of the University of Montpellier. Two type of models are being considered : one for cracks roughly orthogonal to the material fibers, and one for longitudinal cracks, roughly parallel to material fibers.

**Models coupling.** The second step in the modeling phase is to take into account the reduction of fiber strength due to fiber oxidation in structural mechanics simulations. In this coupling, once a description of the fiber weakening due to oxidation is given (see *e.g.* [164]), a material sample can be loaded and a preliminary crack distribution computed using the structural mechanics solver developed in [78]. The mesoscale physicochemical behavior can be simulated with this configuration to get an adjourned variable numerical closure for the structural response of the material. Using this closure, structural mechanics simulations are repeated for the weakened composite. The procedure is repeated until failure. The objective of the PhD of G. Perrot is to be able to reproduce the traction of a single tow of material discussed *e.g.* in [188, 176, 116]. The mid/long term objective is to be able to couple longitudinal and transversal crack models to reproduce the behavior of at least a single composite woven cell. This would allow to test different weaving topologies and bring already the study to the level of a material design tool.

**Numerics.** So far, a standard (continuous) Galerkin discretization of the oxygen diffusion-reaction model has been implemented. The main issues that arise in this type of applications is the variety of time scales related to the very high ratios between liquid polymer/air oxygen diffusion coefficient (of the order of  $10^5$ ), of the (composition dependent) liquid polymer/air viscosities, etc. Even though this type of stiffness can be dealt with by using properly chosen implicit multistep time integration schemes, some form of local time stepping will be needed especially when coupling different models. Ad-hoc finite element discretizations for the full crack averaged model will also be studied.

---

<sup>3</sup>Laboratoire des Composites Thermo-Structuraux

# Publications by the author

- [ABJR11] R. Abgrall, G. Baurin, P. Jacq, and M. Ricchiuto. Some examples of high order simulations parallel of inviscid flows on unstructured and hybrid meshes by residual distribution schemes. *Computers and Fluids*, In Press, Corrected Proof, 2011. doi:10.1016/j.compfluid.2011.05.014.
- [ABSR] R. Abgrall, G. Baurin, D. De Santis, and M. Ricchiuto. Numerical approximation of parabolic problems by means of residual distribution. in preparation, preprint available.
- [ALR08a] R. Abgrall, A. Larat, and M. Ricchiuto. Construction of high order residual distribution schemes. *VKI LS 2003-05, 35<sup>th</sup> Computational Fluid dynamics/ADIGMA Course on very high order discretization methods, von Karman Institute for Fluid Dynamics*, 2008.
- [ALR08b] R. Abgrall, A. Larat, and M. Ricchiuto. Very high order residual distribution schemes for steady flow problems. In *ICCFD5 International Conference on Computational Fluid Dynamics 5*, Seoul, Korea, July 2008.
- [ALR10] R. Abgrall, A. Larat, and M. Ricchiuto. Construction of high-order non upwind distribution schemes. In Norbert Kroll, Heribert Bieler, Herman Deconinck, Vincent Couaillier, Harmen van der Ven, and Kaare Srensen, editors, *ADIGMA - A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*, volume 113 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 107–128. Springer Berlin / Heidelberg, 2010.
- [ALR11] R. Abgrall, A. Larat, and M. Ricchiuto. Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes. *Journal of Computational Physics*, 230(11):4103 – 4136, 2011.
- [ALRT09] R. Abgrall, A. Larat, M. Ricchiuto, and C. Tavé. A simple construction of very high order non-oscillatory compact schemes on unstructured meshes. *Computers & Fluids*, 38(7):1314 – 1323, 2009.
- [AR] R. Abgrall and M. Ricchiuto. Genuinely explicit residual schemes for conservation laws : high order case. in preparation.
- [ARN<sup>+</sup>06] R. Abgrall, M. Ricchiuto, N.Villedieu, C.Tavé, and H.Deconinck. Very high order residual distribution on triangular grids. In *ECCOMAS CFD 2006, European Conference on Computational Fluid Dynamics*, The Netherlands, September 2006.

- [ARTL07] R. Abgrall, M. Ricchiuto, C. Tavé, and A. Larat. Non-oscillatory, very high order residual distribution for steady hyperbolic conservation laws. In *14th Int. Conf. on Finite Elements in Flow Problems*, Santa Fe (NM), USA, 2007.
- [CBR<sup>+</sup>09] C.Y.Kuo, B.Nkonga, M. Ricchiuto, Y.-c.Tai, and B.Bracconnier. Dry granular flows with erosion/deposition process. *ESAIM : PROC*, 28:135–149, 2009.
- [CdSR<sup>+</sup>00] A. Csik, H. de Sterck, M. Ricchiuto, S. Poedts, H. Deconinck, and D. Roose. Explicit and implicit parallel upwind monotone residual distribution solver for the time dependent ideal 2d and 3d magneto-hydrodynamic equations on unstructured grids. In *Second International Conference on Engineering Computational Techniques*, Leuven, Belgium, 2000.
- [CRD02] Á. Csík, M. Ricchiuto, and H. Deconinck. A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws. *J. Comput. Phys*, 179(2):286–312, 2002.
- [CRD03a] A. Csik, M. Ricchiuto, and H. Deconinck. Residual distribution for two-dimensional euler and two-phase flow simulations. In Périaux, Champion, Gagnepain, Pironneau, Stoufflet, and Thomas, editors, *Fluid Dynamics and Aeronautics : New Challenges*, Series of Handbooks on Theory and Engineering Applications of Computational Methods, pages 243–266, 2003. ISBN 84-95999-12-9.
- [CRD03b] Á. Csík, M. Ricchiuto, and H. Deconinck. Space time residual distribution schemes for hyperbolic conservation laws over linear and bilinear elements. *VKI LS 2003-05, 33<sup>rd</sup> Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [CRDP01] Á. Csík, M. Ricchiuto, H. Deconinck, and S. Poedts. Space-time residual distribution schemes for hyperbolic conservation laws. *15th AIAA Computational Fluid Dynamics Conference*, Anaheim, CA, USA, June 2001.
- [CVD<sup>+</sup>11] G. Couegnat, G.L. Vignoles, V. Drean, C. Mulat, W. Ros, G. Perrot, T. Haurat, J. El-Yagoubi, E. Martin, M. Ricchiuto, C. Germain, and M. Cataldi. Virtual material approach to self healing cmcs. In *4th European Conference for Aerospace Sciences (EUCASS)*, St. Petersburg, Russia, July 2011.
- [DPC<sup>+</sup>12] V. Dréan, G. Perrot, G. Couégnat, M. Ricchiuto, and G. L. Vignoles. Image-based 2d numerical modeling of oxide formation in self-healing cmcs. Submitted to the 36th International Conference and Expo on Advanced Ceramics and Composites, 2012.
- [DR07] H. Deconinck and M. Ricchiuto. Residual distribution schemes: foundation and analysis. In E. Stein, R. de Borst, and T.J.R. Hughes, editors, *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Ltd., 2007. DOI: 10.1002/0470091355.ecm054.
- [DRAD11] S. D’Angelo, M. Ricchiuto, R. Abgrall, and H. Deconinck. Generalized framework for adjoint error estimation of PG method in linear problems. Research Report RR-7613, INRIA, 2011.

- [DRD02] J. Dobeš, M. Ricchiuto, and H. Deconinck. Implicit space-time residual distribution for unsteady viscous flow. *16th AIAA Computational Fluid Dynamics Conference*, Florida, USA, June 2002.
- [DRD03a] J. Dobeš, M. Ricchiuto, and H. Deconinck. Implicit space-time residual distribution method for unsteady laminar viscous flow. *VKI LS 2003-05, 33<sup>rd</sup> Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [DRD03c] J. Dobeš, M. Ricchiuto, and H. Deconinck. Implicit space-time residual distribution method for unsteady viscous flow. In *6th Congress on Theoretical and Applied Mechanics*, Ghent, Belgium, 2003.
- [DRD05] J. Dobeš, M. Ricchiuto, and H. Deconinck. Implicit space-time residual distribution method for unsteady laminar viscous flow. *Computers & Fluids*, 34(4-5):593 – 615, 2005.
- [DRV10a] V. Drean, M. Ricchiuto, and G.L. Vignoles. Two-dimensional oxydation modelling of MAC composite materials. Research Report RR-7417, INRIA, 2010.
- [DRV10b] V. Drean, M. Ricchiuto, and G.L. Vignoles. Two-dimensional oxydation modelling of MAC composite materials: part II. Research Report RR-7418, INRIA, 2010.
- [HR09] M. Hubbard and M. Ricchiuto. An unconditionally positive scheme for hyperbolic conservation laws. In *23rd Biennial Conf. on Num. Analysis*, University of Strathclyde (UK), June 2009.
- [HR10] M. Hubbard and M. Ricchiuto. Discontinuous space-time residual distribution: route to unconditional positivity and high order of accuracy. In *International Conference on Fluid Dynamics ICFD10*, Reading, UK, April 2010.
- [HR11] M. Hubbard and M. Ricchiuto. Discontinuous upwind residual distribution: A route to unconditional positivity and high order accuracy. *Computers and Fluids*, 46(1):263 – 269, 2011.
- [HRS11] M. Hubbard, M. Ricchiuto, and D. Sarmani. Residual distribution for shallow water flows. In *Numerical Methods for Hyperbolic Equations*, University of Santiago de Compostela, Spain, July 2011.
- [LAR09] A. Larat, R. Abgrall, and M. Ricchiuto. Construction of high order residual distribution schemes for compressible flow problems. In *ICOSAOM09, International Conference on Spectral and High order Methods*, Trondheim, Norway, 2009.
- [LR] A. Larat and M. Ricchiuto. Conservative  $p^k p^m$  space-time residual discretizations for conservation laws : one dimensional case. in preparation.
- [MRAD03] M. Mezine, M. Ricchiuto, R. Abgrall, and H. Deconinck. Monotone and stable residual distribution schemes on prismatic space-time elements for unsteady conservation laws. *VKI LS 2003-05, 33<sup>rd</sup> Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics -*, 2003.

- [QRCD02] T. Quintino, M. Ricchiuto, Á. Csík, and H. Deconinck. Conservative multidimensional upwind residual distribution schemes for arbitrary finite elements. In *ICCFD2 International Conference on Computational Fluid Dynamics 2*, Sidney, Australia, July 2002.
- [RA06] M. Ricchiuto and R. Abgrall. Stable and convergent residual distribution for time-dependent conservation laws. In *ICCFD4 International Conference on Computational Fluid Dynamics 4*, Ghent, Belgium, July 2006.
- [RA10] M. Ricchiuto and R. Abgrall. Explicit runge-kutta residual distribution schemes for time dependent problems: Second order case. *Journal of Computational Physics*, 229(16):5653 – 5691, 2010.
- [RAA09] M. Ricchiuto, R. Abgrall, and A.Larat. Non-oscillatory high-order residual distribution schemes for the euler equations. In *First symposium on Current and New Trends in Scientific Computing*, Santiago (Chile), October 2009.
- [RAD03] M. Ricchiuto, R. Abgrall, and H. Deconinck. Construction of very high order residual distribution schemes for unsteady advection: preliminary results. *VKI LS 2003-05, 33<sup>rd</sup> Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [RAD07] M. Ricchiuto, R. Abgrall, and H. Deconinck. Application of conservative residual distribution schemes to the solution of the shallow water equations on unstructured meshes,. *J. Comput. Phys.*, 222:287–331, 2007.
- [RB09a] M. Ricchiuto and A. Bollermann. Accuracy of stabilized residual distribution for shallow water flows including dry beds. In E.Tadmor, J.G.Liu, and A.Tzavaras, editors, *HYP08: 12th international conference on hyperbolic problems : theory, numerics, applications*, volume 67(2). AMS, American Mathematical Society, 2009.
- [RB09b] M. Ricchiuto and A. Bollermann. Stabilized residual distribution for shallow water simulations. *J. Comput. Phys*, 228(4):1071–1115, 2009.
- [RCD01] M. Ricchiuto, Á. Csík, and H. Deconinck. Space-time residual distribution and application to unsteady two-phase computations on unstructured meshes. VKI Report VKI PR2001-23, June 2001.
- [RCD04] M. Ricchiuto, Á. Csík, and H. Deconinck. Conservative residual distribution schemes for general unsteady systems of conservation laws. In *ICCFD3 International Conference on Computational Fluid Dynamics 3*, Toronto, Canada, July 2004.
- [RCD05] M. Ricchiuto, Á. Csík, and H. Deconinck. Residual distribution for general time dependent conservation laws. *J. Comput. Phys*, 209(1):249–289, 2005.
- [RCGA12] M. Ricchiuto, P.M. Congedo, G. Geraci, and R. Abgrall. Numerical methods for a reliable numerical prediction of long water-wave phenomena. *14th AIAA Non-Deterministic Approaches Conference*, Honolulu - Hawaii (USA), April 2012.

- [RD99] M. Ricchiuto and H. Deconinck. Time accurate solution of hyperbolic partial differential equations using fct and residual distribution. VKI report VKI SR1999-33, September 1999.
- [RD00] M. Ricchiuto and H. Deconinck. Two-phase flow computations using a two-fluid model and fluctuation splitting schemes. VKI report VKI SR2000-13, June 2000.
- [RD02] M. Ricchiuto and H. Deconinck. Multidimensional upwinding and source terms in inhomogeneous conservation laws: the scalar case. In R. Herbin and D. Kroner, editors, *Finite Volumes for Complex Applications III*. HERMES Science Publishing Ltd, London, 2002.
- [Ric] M. Ricchiuto. An explicit residual approach for shallow water simulations. in preparation, preprint available.
- [Ric05] M. Ricchiuto. Construction and analysis of compact residual discretizations for conservation laws on unstructured meshes. PhD Thesis, Aerospace and aeronautics Department von Karman Institute for Fluid Dynamics and Université Libre de Bruxelles, 2005.
- [Ric09a] M. Ricchiuto. On the c-property and generalized c-property of residual distribution in shallow water simulations. In *First symposium on Current and New Trends in Scientific Computing*, Santiago (Chile), October 2009.
- [Ric09b] M. Ricchiuto. Stabilized residual distribution for shallow water simulations. In *1st international conference on the numerical approximation of hyperbolic systems with source terms, and applications*, Castro-Urdiales, Spain, 2009.
- [Ric11] M. Ricchiuto. On the c-property and generalized c-property of residual distribution for the shallow water equations. *Journal of Scientific Computing*, 48:304–318, 2011.
- [RRWD03] M. Ricchiuto, D.T. Rubino, J.A.S. Witteveen, and H. Deconinck. A residual distributive approach for one-dimensional two-fluid models and its relation with godunov finite volume schemes. In *Proc. of the International Workshop on Advanced Numerical Methods for Multidimensional Simulation of Two-Phase Flow*, Garching, Germany, 2003.
- [RVAD05] M. Ricchiuto, N. Villedieu, R. Abgrall, and H. Deconinck. High order residual distribution schemes: discontinuity capturing crosswind dissipation and extension to advection diffusion. *VKI LS 06-01, 34<sup>rd</sup> CFD course, von Karman Insitute for Fluid Dynamics*, 2005. ISBN 2-930389-63-X.
- [RVAD08] M. Ricchiuto, N. Villedieu, R. Abgrall, and H. Deconinck. On uniformly high-order accurate residual distribution schemes for advection-diffusion. *Journal of Computational and Applied Mathematics*, 215(2):547 – 556, 2008.
- [SFW<sup>+</sup>05] H. Staedtke, G. Franchello, B. Worth, U. Graf, P. Romstedt, A. Kumbaro, J. Garcia-Cascales, H. Paillère, H. Deconinck, M. Ricchiuto, B. Smith, F. De Cachard, E.F. Toro, E. Romenski, and S. Mimouni. Advanced three-dimensional two-phase flow simulation tools for application to reactor safety (astar). *Nuclear Engineering and Design*, 235(2-4):379 – 400, 2005.



- [SHR] D. Sarmani, M. Hubbard, and M. Ricchiuto. Space time residual distribution for shallow water flows. in preparation.
- [VQRD11] N. Villedieu, T. Quintino, M. Ricchiuto, and H. Deconinck. Third order residual distribution schemes for the navier-stokes equations. *Journal of Computational Physics*, 230(11):4301 – 4315, 2011.
- [VRD00] E. Valero, M. Ricchiuto, and G. Degrez. Two-phase flow computations using a two-fluid model and fluctuation splitting. *Trends in Numerical and Physical Modeling for Industrial Two-Phase Flows*, Cargese, France, September 2000.
- [VRD06] N. Villedieu, M. Ricchiuto, and H. Deconinck. High order residual distribution : discontinuity capturing, crosswind dissipation and diffusion. In *ICCFD4 International Conference on Computational Fluid Dynamics 4*, Ghent, Belgium, July 2006.
- [VRQD11] M. Vymazal, M. Ricchiuto, T. Quintino, and H. Deconinck. Variable distribution coefficient residual distribution. In *6th International Conference on Finite Elements in Flow Problems (FEF 2011)*, Munich, Germany, March 2011.

# Bibliography

- [1] R. Abgrall. Toward the ultimate conservative scheme : Following the quest. *J. Comput. Phys*, 167(2):277–315, 2001.
- [2] R. Abgrall. Essentially non oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys*, 214(2):773–808, 2006.
- [3] R. Abgrall. Discontinuous fluctuation distribution. *Adv. Appl. Math. Mech*, 2(1):32–44, 2010.
- [4] R. Abgrall, N. Andrianov, and M. Mezzine. Towards very high-order accurate schemes for unsteady convection problems on unstructured meshes. *Int. J. Numer. Methods Fluids*, 47(8-9):679–691, 2005.
- [5] R. Abgrall and T.J. Barth. New results for residual distribution schemes. In *Toro, E. F. (ed.), Godunov methods. Theory and applications. International conference, Oxford, GB, October 1999. New York, NY: Kluwer Academic/ Plenum Publishers. 27-43.* 2001.
- [6] R. Abgrall and T.J. Barth. Residual distribution schemes for conservation laws via adaptive quadrature. *SIAM J. Sci. Comput.*, 24(3):732–769, 2002.
- [7] R. Abgrall and F. Marpeau. Residual distribution schemes on quadrilateral meshes. *J. Sci. Comput.*, 30(1), 2007.
- [8] R. Abgrall, K. Mer, and B. Nkonga. A Lax–Wendroff type theorem for residual schemes. In M. Hafeez and J.J. Chattot, editors, *Innovative methods for numerical solutions of partial differential equations*, pages 243–266. World Scientific, 2002.
- [9] R. Abgrall and M. Mezzine. Construction of second-order accurate monotone and stable residual distribution schemes for unsteady flow problems. *J. Comput. Phys.*, 188:16–55, 2003.
- [10] R. Abgrall and M. Mezzine. Construction of second-order accurate monotone and stable residual distribution schemes for steady flow problems. *J. Comput. Phys.*, 195:474–507, 2004.
- [11] R. Abgrall and P.L. Roe. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.*, 19(3):3–36, 2003.
- [12] R. Abgrall and C.-W. Shu. Development of residual distribution schemes for the discontinuous galerkin method: the scalar case with linear elements. *Communications in Computational Physics*, 5(2-4):376–390, 2009.

- [13] R. Abgrall and J. Treflik. An example of high order residual distribution scheme using non-lagrange elements. *Journal of Scientific Computing*, 45:3–25, 2010.
- [14] A. Angrand. Des moteurs plus légers et moins bruyants. [http://www.safran-group.com/IMG/pdf/mag1\\_complet-2.pdf](http://www.safran-group.com/IMG/pdf/mag1_complet-2.pdf) and <http://www.safran-group.com/site-safran/presse-et-medias/safran-magazine/>, June 2007.
- [15] A. Angrand. Leap-x, a trailblazer for tomorrow's aero-engines. [http://www.safran-group.com/IMG/pdf/mag5\\_complet.pdf](http://www.safran-group.com/IMG/pdf/mag5_complet.pdf) and <http://www.safran-group.com/site-safran/presse-et-medias/safran-magazine/>, February 2009.
- [16] N. Aslan. MHD-A: A fluctuation splitting wave model for planar magnetohydrodynamics. *Journal of Computational Physics*, 153(2):437–466, 1999.
- [17] N. Aslan. A visual fluctuation splitting scheme for magnetohydrodynamics with a new sonic fix and euler limit. *Journal of Computational Physics*, 197(1):1–27, 2004.
- [18] N. Aslan and M. Mond. A numerical scheme for ionizing shock waves. *Journal of Computational Physics*, 210(2):401–420, 2005.
- [19] E. Audusse and M.-O. Bristeau. A well-balanced positivity preserving second-order scheme for shallow water flows on unstructured meshes. *Journal of Computational Physics*, 206(1):311 – 333, 2005.
- [20] T.J. Barth. An energy look at the N scheme. Working notes, NASA Ames research center, CA, USA, 1996.
- [21] T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In Kröner, Ohlberger, and Rohde, editors, *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, volume 5 of *Lecture Notes in Computational Science and Engineering*, pages 195–285. Springer-Verlag, Heidelberg, 1998.
- [22] T.J. Barth. Numerical methods for conservation laws on structured and unstructured meshes. *VKI LS 2003-05, 33<sup>rd</sup> Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [23] T.J. Barth and D.C Jespersen. The design and application of upwind schemes on unstructured meshes. AIAA paper 89-0355, January 1989. 27th AIAA Aerospace Sciences Meeting, Reno, Nevada (USA).
- [24] J. Bastin and G. Rogé. A multidimensional fluctuation splitting scheme for the three dimensional euler equations. *M2AN*, 33(6), 1999.
- [25] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.
- [26] A. Bermudez and M.E. Vazquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049 – 1071, 1994.

- [27] C. Bernardi, T. Chacón Rebollo, and M. Restelli. A posteriori analysis of a positive streamwise invariant discretization of a convection-diffusion equation. <http://hal.archives-ouvertes.fr/hal-00519691/en/>, submitted to *J.Sci.Comp.*, 2010.
- [28] P. Birken and A. Jameson. On nonlinear preconditioners in newton-krylov methods for unsteady flows. *International Journal for Numerical Methods in Fluids*, 62(5):565–573, 2010.
- [29] P.B. Bochev, M.D. Gunzburger, and J.N. Shadid. On inf-sup stabilized finite element methods for transient problems. *Computer Methods in Applied Mechanics and Engineering*, 193(15-16):1471 – 1489, 2004.
- [30] P.B. Bochev, M.D. Gunzburger, and J.N. Shadid. Stability of the supg finite element method for transient advection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 193(23-26):2301 – 2323, 2004.
- [31] P.B. Bochev, T.J.R. Hughes, and G. Scovazzi. A multiscale discontinuous galerkin method. In Ivan Lirkov, Svetozar Margenov, and Jerzy Wasniewski, editors, *Large-Scale Scientific Computing*, volume 3743 of *Lecture Notes in Computer Science*, pages 84–93. Springer Berlin / Heidelberg, 2006.
- [32] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *R.A.I.R.O. Analyse Numérique*, 12:237–254, 1978.
- [33] A. Bonfiglioli. Multidimensional residual distribution schemes for the pseudo-compressible euler and navier-stokes equations on unstructured meshes. In C H Bruneau, editor, *Lecture Notes in Physics*, pages 254–259. Springer-Verlag, Berlin Heidelberg, 1998.
- [34] A. Bonfiglioli. Fluctuation splitting schemes for the compressible and incompressible euler and navier-stokes equations. *International Journal of Computational Fluid Dynamics*, 14:21–39, 2000.
- [35] A. Bonfiglioli. Hyperbolic-elliptic splitting for the pseudo-compressible euler equations. In E F Toro, editor, *Godunov methods : theory and applications.*, pages 135–140. Kluwer Academic/Plenum Publishers, New-York, 2001.
- [36] A. Bonfiglioli and H. Deconinck. Multidimensional upwind residual distribution schemes for the 3d euler equations. In H Deconinck and B Koren, editors, *Notes on Numerical Fluid Mechanics*, pages 141–185. Vieweg-Verlag, Braunschweig, Germany, 1997.
- [37] P. Bonneton, E. Barthelemy, F. Chazel, R. Cienfuegos, D. Lannes, F. Marche, and M. Tissier. Recent advances in serre-green naghdi modelling for wave transformation, breaking and runup processes. *European Journal of Mechanics - B/Fluids*, In Press, Corrected Proof:–, 2011.
- [38] D.L. Book, J.P. Boris, and K. Hain. Flux-corrected transport ii: Generalizations of the method. *Journal of Computational Physics*, 18(3):248 – 283, 1975.
- [39] J.P. Boris and D.L. Book. Flux-corrected transport. i. shasta, a fluid transport algorithm that works. *Journal of Computational Physics*, 11(1):38 – 69, 1973.

- [40] J.P. Boris and D.L. Book. Flux-corrected transport. iii. minimal-error fct algorithms. *Journal of Computational Physics*, 20(4):397 – 431, 1976.
- [41] L. Bortels. *The Multi-Dimensional Upwinding method as a simulation tool for the analysis of Multi-Ion Electrolytes controlled by Diffusion, Convection and Migration*. PhD thesis, Vrije Universiteit Brussel, 1996.
- [42] L. Bortels, J. Deconinck, and B. Van Den Bossche. The multi-dimensional upwinding method as a new simulation tool for the analysis of multi-ion electrolytes controlled by diffusion, convection and migration. part 1. steady state analysis of a parallel plane flow channel. *Journal of Electroanalytical Chemistry*, 404(1):15 – 26, 1996.
- [43] C. Bourdarias, T. Ngom, and M. Gisclon. Derivation and numerical study of a new viscous shallow water bi-fluid model. *International Journal of Pure and Applied Mathematics*, 68, 2011.
- [44] M.J. Briggs, C.E. Synolakis, G.S. Harkins, and D.R. Green. Laboratory experiments of tsunami runup on a circular island. *Pure and Applied Geophysics*, 144:569–593, 1995.
- [45] P. Brufau and P. Garcia-Navarro. Unsteady free surface flow simulation over complex topography with a multidimensional upwind technique. *Journal of Computational Physics*, 186(2):503 – 526, 2003.
- [46] P. Brufau, P. Garcia-Navarro, and M.E. Vazquez-Cendon. Zero mass error using unsteady wetting-drying conditions in shallow flows over dry irregular topography. *Int. J. Numer. Meth. Fluids*, 45:1047–1082, 2004.
- [47] P. Brufau, M.E. Vazquez-Cendon, and P. Garcia-Navarro. A numerical model for the flooding and drying of irregular domains. *Int. J. Numer. Meth. Fluids*, 39:247–275, 2002.
- [48] E. Burman. Consistent supg-method for transient transport problems: Stability and convergence. *Computer Methods in Applied Mechanics and Engineering*, 199(17-20):1114 – 1123, 2010.
- [49] E. Burman, A. Ern, and M.A. Fernandez. Explicit runge-kutta schemes and finite elements with symmetric stabilization for first-order linear pde systems. *SIAM J. Numer. Anal.*, 48(6):2019–2042, 2010.
- [50] E. Burman and M.A. Fernandez. Finite element methods with symmetric stabilization for the transient convection-diffusion-reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 198(33-36):2508 – 2519, 2009.
- [51] V. Caleffi, A. Valliani, and A. Zanni. Finite volume method for simulating extreme flood events in natural channels. *J.of Hydraulic Research*, 41(2):167–177, 2003.
- [52] D. Caraeni and L. Fuchs. Compact third-order multidimensional upwind scheme for Navier-Stokes simulations. *Theoretical and Computational Fluid Dynamics*, 15:373–401, 2002.
- [53] D. Caraeni and L. Fuchs. Compact third-order multidimensional upwind discretization for steady and unsteady flow simulations. *Computers and Fluids*, 34(4-5):419–441, 2005.

- [54] D.A. Caraeni. *Development of a Multidimensional Upwind Residual Distribution Solver for Large Eddy Simulation of Industrial Turbulent Flows*. PhD thesis, Lund Institute of Technology, 2000.
- [55] J.-C. Carette, H. Deconinck, H. Paillère, and P.L. Roe. Multidimensional upwinding: its relation to finite elements. *International Journal for Numerical Methods in Fluids*, 20:935–955, 1995.
- [56] J. Casado-Diaz, T. Chacon Rebollo, V. Girault, M. Gomez Mormol, and F. Murat. Psi solution of convection-diffusion equations with data in  $l^1$ . In Karl Kunisch, Gunther Of, and Olaf Steinbach, editors, *Numerical Mathematics and Advanced Applications*, pages 233–240. Springer Berlin Heidelberg, 2008.
- [57] M.J. Castro, A.M. Ferreiro Ferreiro, J.A. Garcia-Rodriguez, J.M. Gonzalez-Vida, J. Macias, C. Pares, and M.E. Vazquez-Cendon. The numerical treatment of wet/dry fronts in shallow flows: application to one-layer and two-layer systems. *Mathematical and Computer Modelling*, 42(3-4):419 – 439, 2005.
- [58] M.J. Castro, J. Gonzalez-Vida, and C. Pares. Numerical treatment of wet/dry fronts in shallow flows with a roe scheme. *Mathematical Models and Methods in Applied Sciences*, 16(6):897–931, 2006.
- [59] L. Cea and M.E. Vazquez-Cendon. Unstructured finite volume discretization of bed friction and convective flux in solute transport models linked to the shallow water equations. *J. Comput. Phys.*, 2011. to appear.
- [60] F. Chalot and P.-E. Normand. Higher-order stabilized finite elements in an industrial navier-stokes code. In Norbert Kroll, Heribert Bieler, Herman Deconinck, Vincent Couaillier, Harmen van der Ven, and Kaare Srensen, editors, *ADIGMA - A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*, volume 113 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 145–165. Springer Berlin / Heidelberg, 2010.
- [61] C.-S. Chou and C.-W. Shu. High order residual distribution conservative finite difference weno schemes for steady state problems on non-smooth meshes. *J. Comp. Phys.*, 214(3):698–724, 2006.
- [62] C.-S. Chou and C.-W. Shu. High order residual distribution conservative finite difference weno schemes for convection-diffusion steady state problems on non-smooth meshes. *Journal of Computational Physics*, 224(2):992 – 1020, 2007.
- [63] P.G. Ciarlet and P.A. Raviart. General lagrange and hermite interpolation in  $\mathbb{R}^n$  with applications to finite element methods. *Arch. Ration. Mech. Anal.*, 46:177–199, 1972.
- [64] C. Cluzel, E. Baranger, P. Ladevèze, and A. Mouret. Mechanical behaviour and lifetime modelling of self-healing ceramic-matrix composites subjected to thermomechanical loading in air. *Composites Part A: Applied Science and Manufacturing*, 40(8):976–984, 2009.
- [65] B. Cockburn. Discontinuous galerkin methods for convection-dominated problems. In T.J. Barth and H. Deconinck, editors, *High-Order ENO and WENO schemes for computational fluid dynamics*, volume 9 of *Lecture Notes in Computational Science and Engineering*, pages 69–224. Springer-Verlag, Heidelberg, 1999.

- [66] B. Cockburn, S. Hou, and C.-W. Shu. The runge-kutta local projection discontinuous galerkin finite element method for conservation laws iv : The multidimensional case. *Math.Comp.*, 54(190):545–581, 1990.
- [67] B. Cockburn, G.E. Karniadakis, and C.-W. Shu, editors. *Discontinuous Galerkin methods. Theory, computation and applications*, volume 11 of *Lecture Notes in Computational Science and Engineering*, Heidelberg, 2000. Springer-Verlag.
- [68] B. Cockburn and S.-Y. Lin. Tvb runge-kutta local projection discontinuous galerkin finite element method for conservation laws iii : One-dimensional systems. *J.Comput.Phys.*, 84:90–113, 1989.
- [69] B. Cockburn and C.-W. Shu. The runge-kutta local projection  $p^1$  discontinuous galerkin finite element method for scalar conservation laws. *IMA Preprint Series*, 388, 1988. University of Minnesota.
- [70] B. Cockburn and C.-W. Shu. Tvb runge-kutta local projection discontinuous galerkin finite element method for conservation laws ii : General framework. *Math.Comp.*, 52(186):411–435, 1989.
- [71] B. Cockburn and C.-W. Shu. The runge-kutta local projection discontinuous galerkin finite element method for conservation laws iv : Multidimensional systems. *J.Comput.Phys.*, 141:199–224, 1998.
- [72] G. Cohen, P. Joly, J.E. Roberts, and N. Tordjman. High order triangular finite elements with mass lumping for the wave equation. *SIAM J. Numer. Anal.*, 38(6):2047–2078, 2001.
- [73] C. Corre and A.Lerat. High-order residual-based compact schemes for advection-diffusion problems. *Computers and Fluids*, 37(5):505 – 519, 2008. Special Issue Dedicated to Professor M.M. Hafez on the Occasion of his 60th Birthday - Special Issue Dedicated to Professor M.M. Hafez on the Occasion of his 60th Birthday.
- [74] C. Corre and X. Du. A residual-based scheme for computing compressible flows on unstructured grids. *Computers and Fluids*, 38(7):1338 – 1347, 2009. Special Issue Dedicated to Professor Alain Lerat on the Occasion of his 60th Birthday.
- [75] C. Corre, F. Falissard, and A. Lerat. High-order residual-based compact schemes for compressible inviscid flows. *Computers and Fluids*, 36(10):1567 – 1582, 2007. Special Issue Dedicated to Professor Michele Napolitano on the Occasion of his 60th Birthday.
- [76] C. Corre, G. Hanss, and A. Lerat. A residual-based compact scheme for the unsteady compressible navierstokes equations. *Computers and Fluids*, 34(4-5):561–580, 2005.
- [77] C. Corre and A. Lerat. A residual-based compact scheme of optimal order for hyperbolic problems. *Computers and Fluids*, 41(1):94 – 102, 2011. Implicit Solutions of Navier-Stokes Equations. Special Issue Dedicated to Drs. W.R. Briley and H. McDonald.
- [78] G. Couégnat, E. Martin, and J. Lamon. 3d multiscale modeling of the mechanical behavior of woven composite materials. *Ceram. Eng. Sci. Procs.*, 31(2):185–194, 2010.

- [79] A. Csík. Upwind residual distribution schemes for general hyperbolic conservation laws and application to ideal magnetohydrodynamics. PhD Thesis, Aerospace and aeronautics Department von Karman Institute for Fluid Dynamics and University of Leuven, 2002.
- [80] Á. Csík and H. Deconinck. Space time residual distribution schemes for hyperbolic conservation laws on unstructured linear finite elements. *International Journal for Numerical Methods in Fluids*, 40:573–581, 2002.
- [81] Á. Csík, H. Deconinck, and S. Poedts. Monotone residual distribution schemes for the ideal magnetohydrodynamics equations on unstructured grids. *Proc. 14th AIAA Computational Fluid Dynamics Conference, Norfolk, Virginia, 28/6/'99-1/7/'99*, 2, 1999. ISBN 1-56347-297-X, 644-656.
- [82] Á. Csík, H. Deconinck, and S. Poedts. Monotone residual distribution schemes for the ideal magnetohydrodynamics equations on unstructured grids. *AIAA Journal*, 39(8):1532–1541, 2001.
- [83] Á. Csík, H. Deconinck, and S. Poedts. Performance comparison of multidimensional upwind residual distribution and dimensionally split finite volume Roe schemes on the steady solution of conservation laws. In R. Herbin and D. Kroner, editors, *Finite Volumes for complex applications III*. HERMES Science Publishing Ltd, London, 2002.
- [84] H. Deconinck and T.J. Barth. Study of third order residual distribution schemes for advective problems. Personal communication.
- [85] H. Deconinck, Ch. Hirsch, and J. Peuteman. Characteristic decomposition methods for the multidimensional euler equations. In *Lecture Notes in Physics*, volume 264. Springer-Verlag, 1986.
- [86] H. Deconinck, P.L. Roe, and R. Struijs. A multidimensional generalization of Roe's difference splitter for the Euler equations. *Computers and Fluids*, 22(2/3):215–222, 1993.
- [87] H. Deconinck, K. Sermeus, and R. Abgrall. Status of multidimensional upwind residual distribution schemes and applications in aeronautics. AIAA paper 2000-2328, June 2000. AIAA CFD Conference, Denver (USA).
- [88] G. Degrez and E. van der Weide. Upwind residual distribution schemes for chemical non-equilibrium flows. 14th AIAA Computational Fluid Dynamics Conference, Norfolk, USA, June 28 - July 1 1999.
- [89] A.I. Delis, M.Kazolea, and N.A.Kampanis. A robust high-resolution finite volume scheme for the simulation of long waves over complex domains. *Int. J. for Numerical Methods in Fluids*, 56:419–452, 2008.
- [90] A.I. Delis and N.Katsaounis. Relaxation schemes for the shallow water equations. *Int. J. for Numerical Methods in Fluids*, 41:695–719, 2003.
- [91] P. De Palma, G. Pascazio, and M. Napolitano. An accurate fluctuation splitting scheme for the unsteady two-dimensional Euler equations. ECCOMAS CFD Conference, 2001, Swansea, Wales, UK, September 2001.



- [92] P. De Palma, G. Pascazio, G. Rossiello, and M. Napolitano. A second-order accurate monotone implicit fluctuation splitting scheme for unsteady problems. *J. Comput. Phys.*, 208(1):1–33, 2005.
- [93] P. De Palma, G. Pascazio, D.T. Rubino, and M. Napolitano. Residual distribution schemes for advection and advection diffusion problems on quadrilateral cells. *J. Comput. Phys.*, 218(1):159–199, 2006.
- [94] J. Dobeš and H. Deconinck. A second-order space-time residual distribution method for solving compressible flow on moving meshes, January 2005. 43rd AIAA Aerospace Sciences Meeting, Reno, Nevada (USA).
- [95] J. Dobeš. *Numerical Algorithms for the Computation of Unsteady Compressible Flow over Moving Geometries - Application to Fluid-Structure Interaction*. PhD thesis, Université Libre de Bruxelles, 2007.
- [96] J. Dobeš and H. Deconinck. Second order blended multidimensional upwind residual distribution scheme for steady and unsteady computations. *J. Comput. Appl. Math.*, 215(1):378–389, 2006.
- [97] J. Dobeš and H. Deconinck. An ale formulation of the multidimensional residual distribution scheme for computations on moving meshes. In Herman Deconinck and E. Dick, editors, *Computational Fluid Dynamics 2006*, pages 95–100. Springer Berlin Heidelberg, 2009.
- [98] J. Dobeš, J. Furst, H. Deconinck, and J. Fort. Numerical solution of transonic and supersonic 2d and 3d fluidelastic structure interaction problems. In Karl Kunisch, Gunther Of, and Olaf Steinbach, editors, *Numerical Mathematics and Advanced Applications*, pages 539–546. Springer Berlin Heidelberg, 2008.
- [99] J. Donea. A taylorgalerkin method for convective transport problems. *International Journal for Numerical Methods in Engineering*, 20(1):101–119, 1984.
- [100] J. Donea and A. Huerta. *Unsteady Convection Diffusion Problems*, pages 209–264. John Wiley & Sons, Ltd, 2005.
- [101] M. Dumbser, M. Castro, C. Pares, and E.F. Toro. Ader schemes on unstructured meshes for nonconservative hyperbolic systems: Applications to geophysical flows. *Computers & Fluids*, 38(9):1731 – 1748, 2009.
- [102] M. Dumbser, C. Enaux, and E.F. Toro. Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. *Journal of Computational Physics*, 227(8):3971 – 4001, 2008.
- [103] M. Dumbser and C.D. Munz. Arbitrary high order discontinuous galerkin schemes. numerical methods for hyperbolic and kinetic problems. In Goudon & E. Sonnendrucker eds). *IRMA Series in Mathematics and Theoretical Physics*, pages 295–333. EMS Publishing House, 2005.
- [104] M. Dumbser and C.D. Munz. Building blocks for arbitrary high order discontinuous galerkin schemes. *Journal of Scientific Computing*, 27:215–230, 2006. 10.1007/s10915-005-9025-0.

- [105] A. Ern and J.-C. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer, 2004.
- [106] A. Ern, S. Piperno, and K. Djadel. A well-balanced runge–kutta discontinuous galerkin method for the shallow-water equations with flooding and drying. *Int. J. for Numerical Methods in Fluids*, 58(1):1–25, 2008.
- [107] A. Ferrante and H. Deconinck. Solution of the unsteady Euler equations using residual distribution and flux corrected transport. Technical Report VKI-PR 97-08, von Karman Institute for Fluid Dynamics, 1997.
- [108] J.M. Fiard, T.A. Manteuffel, and S.F. McCormick. First order system least squares for convection diffusion problems : numerical results. *SIAM J. Sci. Comp.*, 19(6):1958–1979, 1998.
- [109] U.S. Fjordholm and S. Mishra. Vorticity preserving schemes finite volume schemes for the shallow water equations. *SIAM J. Sci. Comp.*, 33(2):588–611, 2011.
- [110] L.P. Franca, S.L. Frey, and T.J.R. Hughes. Stabilized finite element methods : I. application to the advective-diffusive model. Technical Report RR-1300, INRIA, 1990.
- [111] T. Gallouët, J.-M. Hérard, and N. Seguin. Some approximate godunov schemes to compute shallow-water equations with topography. *Computers & Fluids*, 32(4):479–513, 2003.
- [112] P. Garcia-Navarro, J. Burguete, and R. Aliod. Numerical simulation of runoff over dry beds. *Monografias del Semin. Matem. Garcia de Galdeano*, 27:307–314, 2003.
- [113] P. Garcia-Navarro, M. Hubbard, and A. Priestley. Genuinely multidimensional upwinding for the 2D shallow water equations. *J. Comp. Phys.*, 121(1):79–93, 1995.
- [114] G. Gassner, M. Dumbser, F. Hindenlang, and C.-D. Munz. Explicit one-step time discretizations for discontinuous galerkin and finite volume schemes based on local predictors. *Journal of Computational Physics*, 230(11):4232 – 4247, 2011.
- [115] G. Gassner, F. Lorcher, and C.-D. Munz. A discontinuous galerkin scheme based on space-time expansion ii. viscous flow equations in multi dimensions. *Journal of Scientific Computing*, 34:260–286, 2008.
- [116] W. Gauthier and J. Lamon. Delayed failure of high-strength and high-strength multifilament tows and single filaments at intermediate temperatures (500°c800°c). *J.Am.Ceram.Soc.*, 92(3):702–709, 2009.
- [117] J.-F. Gerbeau and B. Perthame. Derivation of viscous saint-venant system for laminar shallow water ; numerical validation. *Discrete and Continuous Dynamical Systems, Ser. B*, 1(1):89–102, 2001.
- [118] J.B. Goodman and R.J. LeVeque. On the accuracy of stable schemes for 2d scalar conservation laws. *Mathematics of Computation*, 45(171):15–2, 1985.
- [119] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001.

- [120] S. Goujard, J.L. Charvet, J.L. Leluan, F. Abbé, and G. Lamazouade. Matériau composite protégé contre loxydation par une matrice autocatrisante et son procédé de fabrication. French Patent N° 95 03606., 1995.
- [121] J.M. Greenberg and A.Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.*, 33:1–16, 1996.
- [122] J.-L. Guermond and G. Kanschat. Asymptotic analysis of upwind discontinuous galerkin approximation of the radiative transport equation in the diffusive limit. *SIAM Journal of Numerical Analysis*, 48:53–78, 2010.
- [123] J.-L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230(11):4248 – 4267, 2011.
- [124] S.M.J. Guzik and C.P.T. Groth. Comparison of solution accuracy of multidimensional residual distribution and godunov-type finite-volume methods. *International Journal of Computational Fluid Dynamics*, pages 61–83, 2008.
- [125] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comput. Phys.*, 49:151–164, 1983.
- [126] R. Hartmann. Adaptive discontinuous Galerkin methods with shock-capturing for the compressible Navier-Stokes equations. *Int. J. Numer. Meth. Fluids*, 51(9–10):1131–1156, 2006.
- [127] R. Hartmann. Adjoint consistency analysis of discontinuous Galerkin discretizations. *SIAM J. Numer. Anal.*, 45(6):2671–2696, 2007.
- [128] R. Hartmann. Error estimation and adjoint based refinement for an adjoint consistent DG discretization of the compressible Euler equations. *Int. J. Computing Science and Mathematics*, 1(2–4):207–220, 2007.
- [129] R. Hartmann, J. Held, and T. Leicht. Adjoint-based error estimation and adaptive mesh refinement for the RANS and  $k$ - $\omega$  turbulence model equations. *J. Comput. Phys.*, 230(11):4268–4284, 2011.
- [130] G. Hauke. A symmetric formulation for computing transient shallow water flows. *Computer Methods in Applied Mechanics and Engineering*, 163(1-4):111–122, 1998.
- [131] G. Hauke and M.H. Doweidar. Fourier analysis of semi-discrete and spacetime stabilized methods for the advectivediffusivereactive equation: I. SUPG. *Comp. Meth. Appl. Mech. Engrg.*, 194(1):45–81, 2005.
- [132] G. Hauke and M.H. Doweidar. Fourier analysis of semi-discrete and spacetime stabilized methods for the advectivediffusivereactive equation: II. SGS. *Comp. Meth. Appl. Mech. Engrg.*, 194(6-8):691–724, 2005.
- [133] G. Hauke and M.H. Doweidar. Fourier analysis of semi-discrete and spacetime stabilized methods for the advectivediffusivereactive equation: III. SGS/GSGS. *Comp. Meth. Appl. Mech. Engrg.*, 195(44-47):6158–6176, 2006.
- [134] G. Hauke, A. Landaberea, I. Garmendia, and J. Canales. On the thermodynamics, stability and hierarchy of entropy functions in fluid flow. *Computer Methods in Applied Mechanics and Engineering*, 195(33-36):4473 – 4489, 2006.

- [135] J.C.C. Henriques and L.M.C. Gato. Use of a residual distribution Euler solver to study the occurrence of transonic flow in wells turbine rotor blades. *Comp. Mech.*, 29(3):243–253, 2002.
- [136] J.C.C. Henriques and L.M.C. Gato. A multidimensional upwind matrix distribution scheme for conservative laws. *Computers and Fluids*, 33:755–769, 2004.
- [137] C. Hirsch. ELSEVIER, Butterworth Heinemann, 2007.
- [138] L.C. Huang. Pseudo-unsteady difference schemes for discontinuous solutions of steady-state one dimensional fluid dynamics problems. *J. Comput. Phys.*, 42:195–211, 1981.
- [139] M. Hubbard. Non-oscillatory third order fluctuation splitting schemes for steady scalar conservation laws. *Journal of Computational Physics*, 222(2):740 – 768, 2007.
- [140] M. Hubbard. Discontinuous fluctuation distribution. *Journal of Computational Physics*, 227(24):10125 – 10147, 2008.
- [141] M. Hubbard and M.J. Baines. Conservative multidimensional upwinding for the steady two-dimensional shallow-water equations. *J. Comput. Phys.*, 138:419–448, 1997.
- [142] M. Hubbard and N. Dodd. A 2d numerical model of wave run-up and overtopping. *Coastal Engineering*, 47(1):1–26, 2002.
- [143] M. Hubbard and P. Garcia-Navarro. Flux difference splitting and the balancing of source terms and flux gradients. *J. Comp. Phys.*, 165(1):89–125, 2000.
- [144] M. Hubbard and A.L. Laird. High order fluctuation splitting schemes for time-dependent advection on unstructured grids. *Comput. Fluids.*, 34(4/5):443–459, 2005.
- [145] M. Hubbard and P.L. Roe. Compact high resolution algorithms for time dependent advection problems on unstructured grids. *Int. J. Numer. Methods Fluids*, 33(5):711–736, 2000.
- [146] T.J.R. Hughes and A. Brook. Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comp. Meth. Appl. Mech. Engrg.*, 32:199–259, 1982.
- [147] T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for CFD I: symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comp. Meth. Appl. Mech. Engrg.*, 54:223–234, 1986.
- [148] T.J.R. Hughes and M. Mallet. A new finite element formulation for CFD III: the generalized streamline operator for multidimensional advective-diffusive systems. *Comp. Meth. Appl. Mech. Engrg.*, 58:305–328, 1986.
- [149] T.J.R. Hughes and M. Mallet. A new finite element formulation for CFD IV: a discontinuity-capturing operator for multidimensional advective-diffusive systems. *Comp. Meth. Appl. Mech. Engrg.*, 58:329–336, 1986.
- [150] T.J.R. Hughes, G. Scovazzi, P.B. Bochev, and A. Buffa. A multiscale discontinuous galerkin method with the computational structure of a continuous galerkin method. *Computer Methods in Applied Mechanics and Engineering*, 195(19-22):2761 – 2787, 2006.

- [151] T.J.R. Hughes and T.E. Tezduyar. Development of time-accurate finite element techniques for first order hyperbolic systems with emphasis on the compressible euler equations. *Comp. Meth. Appl. Mech. Engrg.*, 45(1-3):217–284, 1984.
- [152] E. Issman and G. Degrez. A parallel implicit compressible multidimensional upwind euler/navier-stokes solver on unstructured meshes. In *HPCN Europe '96*, pages 599–606, 1996.
- [153] S. Jin and J.-G. Lin. The effects of numerical viscosities I – slowly moving shocks. *J. Comput. Phys.*, 126:373–389, 1996.
- [154] S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on Pure and Applied Mathematics*, pages 235–277, 1995.
- [155] S. Jund. *Méthodes d'éléments finis d'ordre élevé pour la simulations numérique de la propagation d'ondes*. PhD thesis, Université Lous Pasteur, Strasbourg, 2007.
- [156] S. Jund and S. Salmon. Arbitrary high order finite element schemes and high order mass lumping. *Int.J.Appl.Math.Comput.Sci*, 17(3):375–393, 2007.
- [157] H. Kanayama and H. Dan. A finite element scheme for two-layer viscous shallow-water equations. *Japan Journal of Industrial and Applied Mathematics*, 23:163–191, 2006.
- [158] C.M. Klaij, J.J.W. van der Vegt, and H. van der Ven. Pseudo-time stepping methods for space-time discontinuous galerkin discretizations of the compressible navier-stokes equations. *Journal of Computational Physics*, 219(2):622 – 643, 2006.
- [159] D.A. Knoll and D.E. Keyes. Jacobian-free newton-krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193(2):357 – 397, 2004.
- [160] L. Koloszár, N. Villedieu, T. Quintino, P. Rambaud, H. Deconinck, and J. Anthoine. *AIAA Journal*, 49(5):1021–1037, 2011.
- [161] P. Kunthong and L.L. Thompson. An efficient solver for the high-order accurate time-discontinuous galerkin (tdg) method for second-order hyperbolic systems. *Finite Elements in Analysis and Design*, 41(7-8):729 – 762, 2005.
- [162] P. Kunthong and L.L. Thompson. Stabilized time-discontinuous galerkin methods with applications to structural acoustics. November 2006. Proceedings of IMECE2006, 2006 ASME International Mechanical Engineering Congress and Expo - Chicago, Illinois (USA).
- [163] A. Kurganov and E. Tadmor. Solution of two-dimensional Riemann problems without Riemann solvers. *Numerical Methods for Partial Differential Equations*, 18:548–608, 2002.
- [164] P.e Ladevéze and M. Genet. A new approach to the subcritical cracking of ceramic fibers. *Composites Science and Technology*, 70(11):1575 – 1583, 2010.
- [165] D. Lannes and P. Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21, 2009. 016601 doi:10.1063/1.3053183.

- [166] M. Lemou and L. Mieussens. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM J. Sci. Comp.*, 31:334–368, 2008.
- [167] A. Lerat and C. Corre. A residual-based compact scheme for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 170(2):642–675, 2001.
- [168] A. Lerat and C. Corre. Residual-based compact schemes for multidimensional hyperbolic systems of conservation laws. *Computers and Fluids*, 31(4-7):639–661, 2002.
- [169] A. Lerat, F. Falissard, and J. Sidés. Vorticity-preserving schemes for the compressible euler equations. *J. Comput. Phys.*, 225:635–651, 2007.
- [170] R.J. LeVeque. Wave propagation algorithms for multi-dimensional hyperbolic systems. *J. Comput. Phys.*, 131:327–353, 1997.
- [171] R.J. LeVeque. Balancing source terms and flux gradients in high-resolution godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.*, 146(1):346–365, 1998.
- [172] F. Lorcher, G. Gassner, and C.-D. Munz. A discontinuous galerkin scheme based on a spacetime expansion. i. inviscid compressible flow in one space dimension. *Journal of Scientific Computing*, 32:175–199, 2007.
- [173] R.B. Lowrie. Compact higher-order numerical methods for hyperbolic conservation laws. PhD thesis, Aerospace department, University of Michigan, 1996.
- [174] J. Maerz and G. Degrez. Improving time accuracy of residual distribution schemes. Technical Report VKI-PR 96-17, von Karman Institute for Fluid Dynamics, 1996.
- [175] F. Marche. Derivation of a new two-dimensional viscous shallow water model with varying topography, bottom friction and capillary effects. *European Journal of Mechanics - B/Fluids*, 26(1):49 – 63, 2007.
- [176] J. Martinez-Fernandez and G.N. Morscher. Room and elevated temperature tensile properties of single tow hi-nicalon, carbon interphase, cvi sic matrix minicomposites. *Journal of the European Ceramic Society*, 20(14-15):2627 – 2636, 2000.
- [177] G. May and J. Schöberl. Analysis of a spectral difference scheme with flux interpolation on raviart-thomas elements. 2010. Tech. Report AICES-2010/04-8, Aachen Institute for Advanced Study in Computational Engineering Science.
- [178] L. Mesaros. *Multi-dimensional Fluctuation-Splitting Schemes for the Euler equations on unstructured grids*. PhD thesis, University of Michigan, 1995.
- [179] F. Meseguer, E. Valero, C. Martel, J.M. Vega, and I.E. Parra. Unsteady residual distribution schemes for transition prediction. *Aerospace Science and Technology*, 14(8):564 – 574, 2010.
- [180] M. Mezine. *Conception de Schémas Distributifs pour l'aérodynamique stationnaire et instationnaire*. PhD thesis, École doctorale de mathématiques et informatique, Université de Bordeaux I, 2002.

- [181] C. Michler, H. De Sterck, and H. Deconinck. An arbitrary lagrangian eulerian formulation for residual distribution schemes on moving grids. *Computers and Fluids*, 32(1):59 – 71, 2003.
- [182] S. Mishra and E. Tadmor. Constraint preserving schemes using potential-based fluxes. II. genuinely multidimensional systems of conservation laws. *SIAM J. Numer. Anal.*, 49(3):1023–1045, 2011.
- [183] S. Mishra and E. Tadmor. Constraint preserving schemes using potential-based fluxes. II. multidimensional transport equations. *Commun.Comput.Phys*, 9(3):688–710, 2011.
- [184] M.S. Mock. Systems of conservation laws of mixed type. *J. Diff. Eqns.*, 37:70–88, 1980.
- [185] C. Mugler, O. Planchon, J. Patin, S. Weill, N. Silvera, P. Richard, and E. Mouche. Comparison of roughness models to simulate overland flow and tracer transport experiments under simulated rainfall at plot scale. *Journal of Hydrology*, 402(1-2):25 – 40, 2011.
- [186] G. Narbona-Reina, J.D. Zabsonré, E.D. Fernandez-Nieto, and D. Bresch. Un modelo bicapa de tipo shallow water con efectos de viscosidad et fricción. XXI Congreso de Ecuaciones Diferenciales y Aplicaciones - XI Congreso de Matemática Aplicada, September 2009.
- [187] R. Naslain, A. Guette, F. Rebillat, S. LeGallet, F. Lamouroux, L. Filipuzzi, and C. Louchet. Oxidation mechanisms and kinetics of a SiC-matrix composites and their constituents. *Journal of the European Ceramic Society*, 27:377–388, 2007.
- [188] R. Naslain, J. Lamon, R. Paillet, X. Bourrat, A. Guette, and F. Langlais. Micro/minicomposites: a useful approach to the design and development of non-oxide cmcs. *Composites Part A: Applied Science and Manufacturing*, 30(4):537 – 547, 1999.
- [189] I.K. Nikolos and A.I. Delis. An unstructured node-centered finite volume scheme for shallow water flows with wet/dry fronts over complex topography. *Computer Methods in Applied Mechanics and Engineering*, 198(47-48):3723 – 3750, 2009.
- [190] H. Nishikawa. A first-order system approach for diffusion equation. i: Second-order residual-distribution schemes. *Journal of Computational Physics*, 227(1):315 – 352, 2007.
- [191] H. Nishikawa. A first-order system approach for diffusion equation. ii: Unification of advection and diffusion. *Journal of Computational Physics*, 229(11):3989 – 4016, 2010.
- [192] H. Nishikawa. Robust and accurate viscous discretization via upwind scheme i: Basic principle. *Computers & Fluids*, 49(1):62 – 86, 2011.
- [193] H. Nishikawa, M. Rad, and P.L. Roe. A third-order fluctuation splitting scheme that preserves potential flow. 15th AIAA Computational Fluid Dynamics Conference, Anaheim, CA, USA, June 2001.
- [194] H. Nishikawa and B. van Leer. Optimal multigrid convergence by elliptic/hyperbolic splitting. *Journal of Computational Physics*, 190(1):52 – 63, 2003.

- [195] S. Noelle, N. Pankratz, G. Puppo, and J.R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *J. Comput. Phys.*, 213(2):474–499, 2006.
- [196] S. Noelle, Y. Xing, and C.-W. Shu. High order well-balanced finite volume weno schemes for shallow water equation with moving water. *J. Comput. Phys.*, 226:29–58, 2007.
- [197] S. Noelle, Y. Xing, and C.-W. Shu. High order well-balanced schemes. 2009.
- [198] T. Okusanya, D.L. Darmofal, and J. Peraire. Algebraic multigrid for stabilized finite element discretizations of the navierstokes equations. *Computer Methods in Applied Mechanics and Engineering*, 193(33-35):3667 – 3686, 2004.
- [199] H. Paillère. *Multidimensional Upwind residual Discretization Schemes for the Euler and Navier-Stokes Equations on Unstructured Meshes*. PhD thesis, Université Libre de Bruxelles, 1995.
- [200] H. Paillère, C. Corre, and J. Garcia. On the extension of the AUSM+ scheme to compressible two-fluid models. *Computer and Fluids*, 32(6):891–916, 2003.
- [201] H. Paillere and H. Deconinck. Compact cell vertex convection schemes on unstructured meshes. In H Deconinck and B Koren, editors, *Notes on Numerical Fluid Mechanics*, pages 1–50. Vieweg-Verlag, Braunschweig, Germany, 1997.
- [202] H. Paillere and H. Deconinck. Multidimensional upwind residual distribution schemes for the 2d euler equations. In H Deconinck and B Koren, editors, *Notes on Numerical Fluid Mechanics*, pages 51–112. Vieweg-Verlag, Braunschweig, Germany, 1997.
- [203] H. Paillere, H. Deconinck, and P.L. Roe. Conservative upwind residual-distribution schemes based on the steady characteristics of the euler equations. 12th AIAA Computational Fluid Dynamics Conference, San Diego, CA, USA, 1995.
- [204] H. Paillère, H. Deconinck, R. Struijs, P.L. Roe, L.M. Mesaros, and J.-D. Muller. Computations of inviscid compressible flows using fluctuation splitting on triangular meshes. AIAA paper 93-3301, June 1993.
- [205] H. Paillère, G. Degrez, and H. Deconinck. Multidimensional upwind schemes for the shallow-water equations. *International Journal for Numerical Methods in Fluids*, 26:987–1000, 1998.
- [206] L. Quemard, F. Rebillat, A. Guette, H. Tawil, and C. Louchet-Pouillierie. Degradation mechanisms of a SiC fiber reinforced self-healing matrix composites in simulated combustor environment. *Journal of the European Ceramic Society*, 27:377–388, 2007.
- [207] L. Quemard, F. Rebillat, A. Guette, H. Tawil, and C. Louchet-Pouillierie. Self-healing mechanisms of a SiC fiber reinforced multi-layered ceramic matrix composite in high pressure steam environments. *Journal of the European Ceramic Society*, 27:2085–2094, 2007.
- [208] M. Rad. *A residual distribution approach that preserved potential flow*. PhD thesis, Aerospace Engineering - University of Michigan, 2001.



- [209] M. Rad, H. Nishikawa, and P.L. Roe. Some properties of the residual distribution schemes for the euler equations, 2001. First International Conference on Computational Fluid Dynamics - ICCFD.
- [210] M. Rad and P.L. Roe. An Euler code that can compute potential flow. In *Proc. 2nd Int. Symposium on Finite Volumes*. Hermes, 1999.
- [211] T. Chacon Rebollo, M.Gomez Mormol, and G. Narbona Reina. Numerical analysis of the psi solution of advection-diffusion problems through a petrov-galerkin formulation. *M3AS*, 17(11):1905–1936, 2007.
- [212] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. 1973. Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory.
- [213] F. Ringleb. Exacte lösungen de differentialgleichungen einer abadiatischen gasströmung. *ZAMM*, 20:185–198, 1940.
- [214] T.W. Roberts. The behavior of flux difference splitting schemes near slowly moving shock waves. *J. Comput. Phys.*, 90:141–160, 1990.
- [215] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43:357–372, 1981.
- [216] P. L. Roe. Linear advection schemes on triangular meshes. Technical Report CoA 8720, Cranfield Institute of Technology, 1987.
- [217] P. L. Roe. “optimum” upwind advection on a triangular mesh. Technical Report ICASE 90-75, ICASE, 1990.
- [218] P.L. Roe. Fluctuations and signals - a framework for numerical evolution problems. In K.W. Morton and M.J. Baines, editors, *Numerical Methods for Fluids Dynamics*, pages 219–257. Academic Press, 1982.
- [219] P.L. Roe. Characteristics based schemes for the Euler equations. *Annual Review of Fluid Mechanics*, 18:337–365, 1986.
- [220] P.L. Roe. Beyond the Riemann problem, part 1. In M. Y. Hussaini, A. Kumar, & M. D. Salas, editor, *Algorithmic Trends in Computational Fluid Dynamics; The Institute for Computer Applications in Science and Engineering (ICASE)/LaRC Workshop*, pages 341–367, 1993.
- [221] P.L. Roe. Multidimensional upwinding : motivation and concepts. VKI-LS 1994-05, 1994. VKI Lecture series : Computational Fluid Dynamics.
- [222] P.L. Roe and D. Sidilkover. Optimum positive linear schemes for advection in two and three dimensions. *SIAM J. Numer. Anal.*, 29(6):1542–1568, 1992.
- [223] G. Rossiello, P. De Palma, G. Pascazio, and M. Napolitano. Third-order-accurate fluctuation splitting schemes for unsteady hyperbolic problems. *J. Comput. Phys.*, 222(1):332–352, 2007.
- [224] G. Rossiello, P. De Palma, G. Pascazio, and M. Napolitano. Second-order-accurate explicit fluctuation splitting schemes for unsteady problems. *Computers & Fluids*, 38(7):1384 – 1393, 2009.

- [225] J.A. Rossmannith. High-order residual distribution schemes for steady 1d relativistic hydrodynamics. In F. Asakura, editor, *Hyperbolic Problems: Theory, Numerics, and Applications II*, pages 259–266. Yokohama Publishers, 2006.
- [226] T. Schwartzkopff, C.D. Munz, and E.F. Toro. Ader: A high-order approach for linear hyperbolic systems in 2d. *Journal of Scientific Computing*, 17:231–240, 2002. 10.1023/A:1015160900410.
- [227] M. Seaid. Non-oscillatory relaxation methods for the shallow water equations in one and two space dimensions. *Int. J. for Numerical Methods in Fluids*, 46:457–484, 2004.
- [228] K. Sermeus and H. Deconinck. Drag prediction validation of multi-dimensional upwind solver. VKI-LS 2003-02, 2003. VKI Lecture series : CFD-Based Aircraft Drag Prediction and Reduction.
- [229] K. Sermeus and H. Deconinck. Solution of steady euler and navier-stokes equations using residual distribution schemes. VKI-LS 2003-05, 2003. 33rd Computational Fluid Dynamics Course - Novel Methods for Solving Convection Dominated Systems.
- [230] K. Sermeus and H. Deconinck. An entropy fix for multidimensional upwind residual distribution schemes. *Computers and Fluids*, 34(4):617–640, 2005.
- [231] D. Serre. *Systems of conservation laws I - Hyperbolicity, Entropies, Shock waves*. Cambridge University Press, 1999.
- [232] F. Shakib and T.J.R. Hughes. A new finite element formulation for computational fluid dynamics: Ix. fourier analysis of space-time galerkin/least-squares algorithms. *Comp. Meth. Appl. Mech. Engrg.*, 87(1):35–58, 1991.
- [233] C.-W. Shu. High order weighted nonoscillatory schemes for convection dominated problems. *SIAM Review*, 51:82–126, 2009.
- [234] C.-W. Shu. *Discontinuous Galerkin Methods*, pages 661–668. John Wiley & Sons Ltd, 2010. R. Blockley and W. Shyy Editors.
- [235] D. Sidilkover. A genuinely multidimensional upwind scheme and efficient multigrid solver for the compressible euler equations. Technical Report ICASE 94-84, ICASE, 1994.
- [236] D. Sidilkover. Some approaches towards constructing optimally efficient multigrid solvers for the inviscid flow equations. *Computers & Fluids*, 28(4-5):551 – 571, 1999.
- [237] D. Sidilkover and P.L. Roe. Unification of some advection schemes in two dimensions. *Technical Report 95-10, ICASE*, 1995.
- [238] S.P. Spekreijse. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comp.*, 49:135–155, 1987.
- [239] R. Struijs. *A Multi-Dimensional Upwind Discretization Method for the Euler Equations on Unstructured Grids*. PhD thesis, University of Delft, Netherlands, 1994.
- [240] R. Struijs, H. Deconinck, P. De Palma, P.L. Roe, and K.G. Powell. Progress on multidimensional upwind Euler solvers for unstructured grids. AIAA paper 91–1550, 1991.

- [241] R. Struijs, H. Deconinck, and P.L. Roe. Fluctuation splitting schemes for the 2D Euler equations. VKI-LS 1991-01, 1991. Computational Fluid Dynamics.
- [242] C.E. Synolakis, E. Bernard, V. Titov, U. Kanoglu, and F. Gonzalez. Validation and verification of tsunami numerical models. *Pure and Applied Geophysics*, 165:2197–2228, 2008.
- [243] J. Szmelter and P.K. Smolarkiewicz. An edge-based unstructured mesh discretisation in geospherical framework. *Journal of Computational Physics*, 229(13):4980 – 4995, 2010.
- [244] J. Szmelter and P.K. Smolarkiewicz. An edge-based unstructured mesh framework for atmospheric flows. *Computers & Fluids*, 46(1):455 – 460, 2011.
- [245] E. Tadmor. Skew-selfadjoint form for systems of conservation laws. *J. Math. Anal. Appl.*, 103:428–442, 1984.
- [246] E. Tadmor. Entropy functions for symmetric systems of conservation laws. *J. Math. Anal. Appl.*, 122:355–359, 1987.
- [247] T.E. Tezduyar and M. Senga. Stabilization and shock-capturing parameters in SUPG formulation of compressible flows. *Comp. Meth. Appl. Mech. Engrg.*, 195:1621–1632, 2006.
- [248] W.C. Thacker. Some exact solutions to the nonlinear shallow-water wave equations. *J. Fluid Mechanics*, 107:499–508, 1981.
- [249] V.A. Titarev and E.F. Toro. Ader: Arbitrary high order godunov approach. *Journal of Scientific Computing*, 17:609–618, 2002.
- [250] E. Valero, J. de Vicente, and G. Alonso. The application of compact residual distribution schemes to two-phase flow problems. *Computers and Fluids*, 38(10):1950 – 1968, 2009.
- [251] K. van den Abeele, G. Ghorbaniasl, M. Parsani, and C. Lacor. A stability analysis for the spectral volume method on tetrahedral grids. *Journal of Computational Physics*, 228(2):257 – 265, 2009.
- [252] K. van den Abeele and C. Lacor. An accuracy and stability study of the 2d spectral volume method. *Journal of Computational Physics*, 226(1):1007 – 1026, 2007.
- [253] K. van den Abeele, C. Lacor, and Z.J. Wang. On the connection between the spectral volume and the spectral difference method. *Journal of Computational Physics*, 227(2):877 – 885, 2007.
- [254] K. van den Abeele, C. Lacor, and Z.J. Wang. On the stability and accuracy of the spectral difference method. *Journal of Scientific Computing*, 37:162–188, 2008. 10.1007/s10915-008-9201-0.
- [255] E. van der Weide. *Compressible Flow Simulation on Unstructured Grids using Multi-dimensional Upwind Schemes*. PhD thesis, Delft University of Technology, Netherlands, 1998.

- [256] E. van der Weide and H. Deconinck. Positive matrix distribution schemes for hyperbolic systems. In *Computational Fluid Dynamics*, pages 747–753, New York, 1996. Wiley.
- [257] E. van der Weide and H. Deconinck. Matrix distribution schemes for the system of euler equations. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes solvers using multidimensional upwind schemes and multigrid acceleration*, volume 57 of *Notes on Numerical Fluid Dynamics*, pages 113–135. Vieweg, 1997.
- [258] E. van der Weide, H. Deconinck, E. Issmann, and G. Degrez. A parallel implicit multidimensional upwind residual distribution method for the Navier-Stokes equations on unstructured grids. *Comp. Mech.*, 23(2):199–208, 1999.
- [259] B. van Leer. Progress in multi-dimensional upwind differencing. In M. Napolitano and F. Sabetta, editors, *Thirteenth International Conference on Numerical Methods in Fluid Dynamics*, volume 414 of *Lecture Notes in Physics*, pages 1–26. Springer Berlin / Heidelberg, 1993.
- [260] B. van Leer, W.-T. Lee, and P.L. Roe. Characteristic time-stepping or local preconditioning for the euler equations. 10th AIAA Computational Fluid Dynamics Conference, 1991.
- [261] N. Villedieu. High order discretization by residual distribution schemes. PhD Thesis, Aerospace and aeronautics Department von Karman Institute for Fluid Dynamics and Université Libre de Bruxelles, 2009.
- [262] R. von Mises. Dover, 1958. Unabridged republication of the work first published by Academic Press Inc.
- [263] Z.J. Wang. Spectral (finite) volume method for conservation laws on unstructured grids: Basic formulation. *J.Comput.Phys.*, 178:210–251, 2002.
- [264] Z.J. Wang and Y. Liu. Spectral (finite) volume method for conservation laws on unstructured grids ii: Extension to two-dimensional scalar equation. *J.Comput.Phys.*, 179:665–697, 2002.
- [265] Z.J. Wang and Y. Liu. Spectral (finite) volume method for conservation laws on unstructured grids iii: Extension to one-dimensional systems. *J.Sci.Comp.*, 20:137–157, 2004.
- [266] Z.J. Wang, Y. Liu, G. May, and A. Jameson. Spectral difference method for unstructured grids ii: Extension to the euler equations. *Journal of Scientific Computing*, 32:45–71, 2007.
- [267] Z.J. Wang, L. Zhang, and Y. Liu. Spectral (finite) volume method for conservation laws on unstructured grids iv: Extension to two-dimensional euler equations. *J.Comput.Phys.*, 194:716–741, 2004.
- [268] N. Waterson and H. Deconinck. Simulation of low-mach number flow using a fully coupled implicit residual distribution method. ECCOMAS CFD Conference, 2006, TU Delft, The Netherlands, 2006.
- [269] N.P. Waterson. *Simulation of turbulent flow, heat and mass transfer using a residual-distribution approach*. PhD thesis, Delft University of Technology, 2003.

- [270] C.H. Whiting, K.E. Jansen, and S. Dey. Hierarchical basis for stabilized finite element methods for compressible flows. *Computer Methods in Applied Mechanics and Engineering*, 192(47-48):5167 – 5185, 2003.
- [271] J.S. Wong, D.L. Darmofal, and J. Peraire. High-order finite element discretization for the compressible euler and navier-stokes equations. 2001. FDRL TR-01-1, Dept. of Aeronautics & Astronautics, MIT.
- [272] W.A. Wood. Multi-dimensional upwind fluctuation splitting scheme with mesh adaptation for hypersonic viscous flow. Technical Report NASA/TP-2002-211640, NASA Langley Research center, 2002.
- [273] W.A. Wood and W.L. Kleb. Diffusion characteristics of finite volume and fluctuation splitting schemes. *J. Comput. Phys.*, 153:353–377, 1999.
- [274] P.R. Woodward and P. Colella. The numerical simulation of two-dimensional flows with strong shocks. *J. Comput. Phys.*, 54:115–173, 1984.
- [275] L. Wu and D.B. Bogy. Numerical simulation of the slider air bearing problem of hard disk drives by two multidimensional upwind residual distribution schemes over unstructured triangular meshes. *Journal of Computational Physics*, 172(2):640–657, 2001.
- [276] Y. Xing and C.-W. Shu. High-order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *J. Comput. Phys.*, 214(2):567–598, 2006.
- [277] Y. Xing and C.-W. Shu. High-order finite volume weno schemes for the shallow water equations with dry states. *Advances in Water Resources*, 34(8):1026 – 1038, 2011.
- [278] Y. Xing, C.-W. Shu, and S. Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *Journal of Scientific Computing*, 48:339–349, 2011.
- [279] Y. Xing, X. Zhang, and C.-W. Shu. Positivity-preserving high order well-balanced discontinuous galerkin methods for the shallow water equations. *Advances in Water Resources*, 33(12):1476 – 1493, 2010.
- [280] M. Yano and D.L. Darmofal. Bddc preconditioning for high-order galerkin least-squares methods using inexact solvers. *Computer Methods in Applied Mechanics and Engineering*, 199(45-48):2958 – 2969, 2010.
- [281] D.W. Zaide and P.L. Roe. Shock capturing anomalies and the jump conditions in one dimension. 20th AIAA CFD Conference, 2011.
- [282] S.T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335 – 362, 1979.
- [283] X. Zhang and C.-W. Shu. On positivity-preserving high order discontinuous galerkin schemes for compressible euler equations on rectangular meshes. *Journal of Computational Physics*, 229(23):8918 – 8934, 2010.

- [284] X. Zhang and C.-W. Shu. Positivity-preserving high order discontinuous galerkin schemes for compressible euler equations with source terms. *Journal of Computational Physics*, 230(4):1238 – 1248, 2011.
- [285] X. Zhang, Y. Xia, and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high order discontinuous galerkin schemes for conservation laws on triangular meshes. *Journal of Scientific Computing*, pages 1–34, 2011. to appear (online first, doi 10.1007/s10915-011-9472-8).
- [286] O.C. Zienkiewicz. *The finite element method : its basis and fundamentals*. ELSEVIER, Butterworth Heinemann, 2005.