

VON KARMAN INSTITUTE
FOR FLUID DYNAMICS

UNIVERSITÉ LIBRE DE BRUXELLES
FACULTÉ DES SCIENCES APPLIQUÉES

Construction and analysis of compact residual discretizations for conservation laws on unstructured meshes

Mario Ricchiuto
May 2005

Thèse présentée en vue de l'obtention du titre de
Docteur en Sciences Appliquées

Doctoral committee:

Prof.	Abgrall, R.	(Université de Bordeaux I)
Prof.	Beauwens, R.	(Université Libre de Bruxelles)
Prof.	Deconinck, H.	(von Karman Institute for Fluid Dynamics, promotor)
Prof.	Degrez, G.	(Université Libre de Bruxelles, president)
Dr.	Delanaye, M.	(CENAERO)
Prof.	Remacle, J.-F.	(Université Catholique de Louvain)

...alla mia famiglia

Contents

Summary	vii
Acknowledgements	ix
1 Introduction	1
1.1 Motivation	1
1.2 Objectives and scopes of the present work	6
1.2.1 Historical overview and literature survey on \mathcal{RD}	6
1.2.2 Open issues	7
1.2.3 The contribution of this thesis	11
1.3 Structure of the manuscript	12
2 Conservation laws: continuous problem and related stability	15
2.1 The scalar advection equation	16
2.2 Scalar nonlinear conservation laws	18
2.2.1 Weak solutions	19
2.2.2 Conservation	19
2.2.3 Entropy and dissipation	19
2.2.4 Max principle	20
2.3 Linear symmetric hyperbolic systems	21
2.4 Nonlinear systems of conservation laws	21
2.5 Summary	23
3 Discrete approximation: grid, geometry and unknowns	25
3.1 Mesh geometry	26
3.2 Variable and flux approximation	27
3.2.1 Discrete approximation of the unknowns: linear case	27
3.2.2 Discrete approximation of the unknowns: nonlinear case	28
3.3 Flux Jacobians	29
3.3.1 Scalar Jacobians	30
4 Prototype compact discrete approximation for steady advection	31
4.1 Positive cell-vertex schemes on unstructured grids	32
4.1.1 Discrete maximum principle: explicit case	33
4.1.2 Discrete maximum principle: implicit case	35
4.2 Energy stability	39
4.2.1 Fully discrete case	44

4.3	Stability: the inhomogeneous case	45
4.3.1	L^∞ stability: inhomogeneous case	46
4.4	Accuracy and Godunov's theorem	48
4.4.1	The steady case	48
4.4.2	Steady inhomogeneous case	50
4.4.3	Time-dependent computations	51
4.4.4	Linear schemes and Godunov's theorem	53
4.5	Summary	54
5	$\mathcal{RD}/\mathcal{FS}$ schemes for steady advection	55
5.1	$\mathcal{RD}/\mathcal{FS}$: definition and generalities	55
5.1.1	The residual	57
5.1.2	A residual property: linearity preserving schemes	59
5.2	Finite Volume schemes in \mathcal{FS} formalism	60
5.2.1	The upwind \mathcal{FV} scheme: positivity and energy stability	62
5.3	Central schemes and \mathcal{FE}	64
5.3.1	The central \mathcal{RD} scheme and the Galerkin \mathcal{FE} method	64
5.3.2	Petrov-Galerkin schemes in \mathcal{RD} form	65
5.3.2.1	PG schemes: energy stability	65
5.3.3	The Rusanov scheme	67
5.3.3.1	Rv scheme: positivity and energy stability	67
5.4	Multidimensional Upwind schemes	68
5.4.1	The LDA scheme	70
5.4.1.1	LDA scheme: energy stability	71
5.4.2	The N scheme	72
5.4.2.1	N scheme: positivity and energy stability	73
5.4.3	Relations between the N and LDA schemes: dissipation	74
5.4.4	An N scheme for inhomogeneous advection	75
5.5	Nonlinear schemes	75
5.5.1	Blended schemes	76
5.5.1.1	Blended schemes and energy stability	77
5.5.2	The PSI scheme: limited nonlinear schemes	77
5.5.2.1	Well-posedness of the mapping	79
5.5.2.2	Well-posed mappings: a counterexample	81
5.5.2.3	Limited schemes and energy stability	82
5.6	Illustrative examples	84
5.6.1	Rotational advection: homogeneous case	84
5.6.2	Rotational advection with a source term	85
5.7	Summary	88
6	Nonlinear scalar conservation laws: \mathcal{CRD} schemes	91
6.1	Conservative and non-conservative schemes	91
6.2	Conservative \mathcal{RD} formulations: \mathcal{CRD} and \mathcal{QRD}	94
6.2.1	The \mathcal{QRD} formulation	95
6.2.2	The \mathcal{CRD} formulation	97
6.2.2.1	Linearity preserving schemes	97
6.2.2.2	The N scheme	98
6.2.2.3	Rusanov's scheme	101

6.3	Conservative residual distribution and entropy	102
6.3.1	Entropy stability, central \mathcal{RD} schemes and \mathcal{FE}	104
6.3.1.1	Galerkin \mathcal{FE} and centered \mathcal{FS} scheme	104
6.3.1.2	Streamline dissipation	105
6.3.1.3	The Rv scheme	106
6.3.2	Entropy dissipation and \mathcal{MU} schemes	106
6.3.2.1	The LDA scheme	107
6.3.2.2	The N scheme	108
6.3.3	Time integration	111
6.3.3.1	Explicit FE time-integration	111
6.3.3.2	Implicit BE time-integration	112
6.3.3.3	Trapezium time scheme and \mathcal{CN} scheme	112
6.4	Computational examples	113
6.5	Summary	115
7	Time dependent problems: conservative space-time \mathcal{RD}	117
7.1	Time-dependent advection	117
7.1.1	An improved prototype for unsteady simulations	118
7.1.1.1	Finite element schemes with mass-matrix	119
7.1.1.2	A \mathcal{RD} Taylor-Galerkin approach: consistent LW scheme	120
7.1.1.3	\mathcal{FS} schemes for time-dependent computations	123
7.1.2	A space-time framework	124
7.1.2.1	Accuracy of space-time \mathcal{RD}	125
7.1.2.2	The space-time residual	126
7.1.3	Geometry of space-time \mathcal{RD} schemes	128
7.1.4	Space-time \mathcal{MU} schemes	130
7.1.4.1	LDA schemes	132
7.1.4.2	N schemes	132
7.1.4.3	Limited schemes	133
7.1.5	Digression: two-layer schemes	134
7.2	Nonlinear conservation laws	136
7.2.1	Conservative space-time schemes	136
7.2.2	\mathcal{CRD} for time-dependent \mathcal{CL} s	137
7.2.2.1	\mathcal{LP} discretizations: \mathcal{CRD} LDA and ST-LDA schemes .	138
7.2.2.2	“Monotone” discretizations: \mathcal{CRD} N and ST-N schemes	138
7.2.2.3	Nonlinear schemes	139
7.3	Computational examples	140
7.3.1	Accuracy study for linear advection	140
7.3.2	A nonlinear problem	145
7.4	Summary	146
8	Extension to systems	149
8.1	Matrix \mathcal{RD} for linear symmetric systems	150
8.1.1	Matrix LDA scheme	153
8.1.1.1	Energy production of the matrix LDA scheme	154
8.1.2	Matrix N scheme	154
8.1.2.1	Energy stability	155
8.1.3	Nonlinear matrix \mathcal{RD} schemes	156

8.2	Space-time matrix \mathcal{RD}	157
8.2.1	Space-time matrix LDA schemes	159
8.2.2	Space-time matrix N schemes	160
8.3	Nonlinear systems	161
8.3.1	\mathcal{CRD} schemes for steady systems of \mathcal{CL} s	163
8.3.1.1	The \mathcal{CRD} LDA scheme	164
8.3.1.2	The \mathcal{CRD} N scheme	165
8.3.1.3	Nonlinear schemes	166
8.3.2	Space-time \mathcal{CRD} schemes	166
8.3.2.1	Space-time \mathcal{CRD} LDA schemes	167
8.3.2.2	Space-time \mathcal{CRD} N schemes	167
8.3.2.3	Nonlinear schemes	168
8.4	Summary	169
9	Computational details	171
9.1	Steady computations	172
9.2	Time-dependent computations	174
9.2.1	Two-layer schemes	176
10	Evaluation on the Euler equations of a perfect gas	177
10.1	Steady computations	178
10.1.1	Mach 10 blunt body flow	178
10.2	Time-dependent computations	181
10.2.1	Mach 10 moving shock	181
10.2.2	A 2D Riemann problem	185
10.2.3	Double Mach reflection	186
10.2.3.1	Grid refinement with the LN scheme	188
10.2.4	A shock-shock interaction	190
10.2.4.1	Grid refinement with the LN scheme	192
10.2.5	A shock-bubble interaction	194
10.2.6	Two-layer schemes: Mach 3 flow over a step	196
10.2.7	Two-layer schemes: slow shock hitting a wedge	197
10.3	Summary	199
11	A two-phase flow model	201
11.1	Time-dependent computations	203
11.1.1	Moving shocks in air-water mixtures	203
11.1.2	A two-phase 2D Riemann Problem	204
11.2	A shock-bubble interaction	206
11.3	Summary	208
12	Solution of the shallow-water equations	211
12.1	\mathcal{RD} schemes and the lake-at-rest solution	213
12.2	Steady-state computations	215
12.2.1	Super-critical flow over flat bed	215
12.2.2	Trans-critical flow over a smooth hump	218
12.3	Time-dependent computations	218
12.3.1	Break of a circular dam over flat bed	218

12.3.2	Non-symmetric dam break over flat bed	221
12.3.3	Water height perturbation over smooth bed	223
12.3.4	Water height perturbation over non-smooth bed	228
12.4	Summary	231
13	Conclusions and perspectives	233
13.1	Main achievements	234
13.1.1	Compact cell-vertex schemes for scalar advection	234
13.1.2	Fluctuation splitting schemes	235
13.1.3	Residual distribution for nonlinear conservation laws	236
13.1.4	Systems of \mathcal{CL} s: verification	238
13.2	Weaknesses of the methodology proposed	240
13.2.1	Nonlinear schemes and stability	240
13.2.2	On the efficiency	241
13.3	Future perspectives	242
13.3.1	Very high-order schemes	242
13.3.2	Viscous terms and sources	243
13.3.3	Hybrid, adaptive and moving meshes	244
13.3.4	Applications	245
	Bibliography	247

Summary

This thesis presents the construction, the analysis and the verification of compact residual discretizations for the solution of conservation laws on unstructured meshes. The schemes considered belong to the class of residual distribution (\mathcal{RD}) or fluctuation splitting (\mathcal{FS}) schemes. The methodology presented relies on three main elements

1. Construction of compact linear first-order stable schemes for linear hyperbolic PDEs;
2. A *positivity preserving* procedure mapping stable first-order linear schemes onto nonlinear second-order schemes with non-oscillatory shock capturing capabilities;
3. A conservative formulation enabling to extend the schemes to nonlinear \mathcal{CL} s.

These three *design* steps and the underlying theoretical tools are discussed in depth. The nonlinear \mathcal{RD} schemes resulting from this construction are tested on a large set of problems involving the solution of scalar models, and systems of \mathcal{CL} s. This extensive verification fills the gaps left open, where no theoretical analysis can be performed. Results are presented on the Euler equations of a perfect gas, on a two-phase flow model with highly nonlinear thermodynamics and on the shallow-water equations. On irregular grids, the schemes proposed yield quite accurate and stable solutions even on very difficult computations. These results are more accurate than the ones given by \mathcal{FV} and WENO schemes. Moreover, our schemes have a compact nearest-neighbor stencil. This encourages to further develop our approach, toward the design of robust very high-order schemes for complex applications. These schemes would represent a very appealing alternative, both in terms of accuracy and efficiency, to now classical \mathcal{FV} and ENO/WENO discretizations. A better understanding of the dissipation properties of the nonlinear discretizations proposed in the thesis might lead to further improvements in efficiency rendering the schemes very competitive also with respect to very high-order \mathcal{DG} schemes.

Acknowledgments

I want to express my gratitude to my supervisor Herman Deconinck. Through all these years, Herman has been a constant reference, helping me to grow from the scientific and personal point of view. He encouraged and trusted me, giving me the chance to work at the highest possible level. I will always be grateful to have had the possibility of working next to him. I'm sure that our collaboration will go beyond this thesis.

To the other members of my doctoral committee, Prof. Abgrall, Prof. Beauwens, Prof. Degrez, Dr. Delanaye, and Prof. Remacle, thanks for taking the time to review the manuscript, and for showing so much interest in my work. I want to further thank Gérard Degrez for the many discussions we had during these years. I'm particularly indebted with Rémi Abgrall. I cannot but admire Rémi's passion for his work and I will never forget the warm hospitality I always enjoyed in Bordeaux. I'm sure that working with him will be pleasant and rewarding. I'm also grateful to Prof. P.L. Roe and to Dr. T.J. Barth for the useful and stimulating discussions we had in several occasions.

I enjoyed working with Á. Csík and J. Dobeš who have contributed to many developments of my work. When things seemed particularly hard, the discussions with people at VKI have often helped me to find back my motivation.

One of my best friends often says: "*Life is an adventure...*" After years, I'm still puzzled with the question of whether he's been trying to be ironic on life in general... or just making fun of me. Anyway, living in Brussels was some kind of adventure. The amazing number of different people I met made it such. VKI students, EU stagiaires, ERASMUS students, people coming to Brussels for one or half a year internships in various companies or institutions. Parties, drinks, trips, dinners, movies, concerts, and more. While the end of my PhD was approaching, most of them have left and some are still in Brussels. To all of them, I owe the best moments of these years. For those friends who have become a very important part of my life, I'm sure that what we share, shared, and will share in the years to come, is more important and rewarding than reading their name on this page.

Lastly, this thesis is dedicated to my parents and to my sister. They have to be proud, since every small success I'm able to achieve, proves how greatly successful they have been in supporting and loving me.

Rhode Saint Genèse, August 2005

Chapter 1

Introduction

1.1 Motivation

The development of high-order algorithms for the simulation of compressible flows in complex domains and on arbitrary meshes is one of the most important research topics in Computational Fluid Dynamics. The continuous growth of the available computing power allows to increase the complexity of the flow configurations object of the simulations. However, improvements in the efficiency, flexibility and robustness of the numerical algorithms are needed to fully exploit this computational potential.

It is generally agreed that, when dealing with complex geometries and flow patterns, the use of unstructured grids is somewhat mandatory. Compared to structured and multi-block structured grids, the generation of unstructured meshes, or more generally hybrid unstructured/structured meshes, can in fact be highly automated and needs a considerably lower degree of *user-input* and, consequently, time [16]. Moreover, unstructured mesh generation lends itself very naturally to solution-dependent local refinement and adaptation, which are known to improve the simulation output, at the same time reducing the number of elements/nodes needed to achieve a fixed level of accuracy [16, 17, 22]. As a consequence, the design of new numerical algorithms for the simulation of compressible flow is largely oriented to formulations well suited for unstructured grids (see *e.g.* the volumes [22, 21]).

An abstract model for the fluid-mechanics equations is given by a so-called *Conservation Law* (\mathcal{CL}): a Partial Differential Equation (PDE) stating the conservation of some unknowns over a given region of space and time. The design of new numerical schemes for compressible flow simulations often starts with the study of simple \mathcal{CL} s for which one has more information on the properties of the exact solution. It is generally accepted that state of the art numerical methods for conservation laws on unstructured grids are not entirely satisfactory. The need of more flexible, accurate and robust

solution algorithms for the analysis of large and complex systems is what drives the development of new techniques. Accuracy, robustness and efficiency requirements lead to the following *design constraints*:

Accuracy It should be possible to increase the accuracy of the approximation in a relatively simple way, without introducing expensive reconstruction steps. Moreover, due to the fact that unstructured grids can be quite irregular (especially in 3D), the accuracy of the method should be as insensitive as possible to the regularity of the mesh;

Stability Conservation laws admit *weak solutions* containing discontinuities. These solutions are piece-wise smooth without strong oscillations in correspondence of the singularities. The numerical method must be able to handle discontinuities without polluting the solution with spurious oscillations. Additionally, weak solutions of \mathcal{CL} s also verify additional constraints imposed by the existence of a (vanishing) dissipative mechanism¹. This gives an additional stability requirement for the numerical method. Ideally, the stability of the scheme (non-oscillatory character and energy/entropy stability) should be *parameter free*, that is, it should not depend on constants which are difficult to optimize in a general way;

Efficiency A numerical method should allow a fast and efficient implementation, particularly on parallel platforms. From this point of view, the main requirements are simplicity and *compactness*. A compact method is one that, to compute the value of the unknowns in a certain mesh location, only uses information contained in the closest grid-entities. In parallel implementations, this allows to minimize the overhead due to inter-processor communication. Compactness is equivalent to the *locality* of the discretization procedure.

On unstructured meshes, state of the art Finite Volume (\mathcal{FV}) schemes are accepted to have strong deficiencies as far as accuracy and efficiency are concerned. This is related to several factors. First of all, in multiple dimensions most \mathcal{FV} schemes are designed by applying their onedimensional formulation along particular mesh directions (edges, edge normals, etc...). This often reduces dramatically the accuracy on irregular meshes. Moreover, the construction of second or higher-order schemes necessitates the local reconstruction of polynomials of the proper degree. This renders the schemes non-compact, hence less efficient. Even though there have been attempts to design truly multidimensional finite volume schemes (see [111, 107] and references therein for example) and improved high order \mathcal{FV} schemes for unstructured meshes [24, 23, 20, 25], the main deficiencies remain. These deficiencies are not cured by the very high-order extensions obtained using the ENO/WENO philosophy (see the reviews [157, 158] and references therein), which are based on even more complex polynomial reconstructions.

A more promising approach is the one at the basis of the so-called *residual-based* discretizations. Even though different in spirit, all residual-based schemes can be seen as some weighted-residual approximation of the conservation law. The advantage of this approach is that it can reproduce exactly solutions in function spaces determined

¹The entropy inequality implied by the second principle of thermodynamics is an example

by the type of *interpolation* used for the unknowns on the mesh. Since this is true (almost) independently on the regularity of the grid, these methods are particularly well suited to work on unstructured meshes. Moreover, the accuracy of the schemes can be increased just by improving the approximation of the unknowns. However, differently from the ENO/WENO approach, these approximations *are not reconstructed* but *defined a-priori* on the elements of the grid. This makes residual methods very compact and efficiently parallelizable. Most residual schemes, in fact, compute the value of the solution in a given location of the mesh only using the information stored in immediately adjacent mesh entities.

Examples of residual methods are stabilized Finite Element (\mathcal{FE}) schemes [94, 97, 95, 96, 102, 103, 166], Discontinuous Galerkin (\mathcal{DG}) schemes (see [42] and references therein), the Residual-Based Compact (RBC) schemes of [109] and the Residual Distribution (\mathcal{RD}) schemes (see [55, 9] and references therein). The application of the RBC method is currently limited to structured meshes, even though promising results on unstructured grids exist [43]. The \mathcal{DG} schemes have shown impressive results. Being based on a stabilized Galerkin approach, as stabilized \mathcal{FE} schemes, they have well defined local energy (and entropy) stability properties, which can be easily proved. However, the stabilization mechanism used in \mathcal{DG} is based on \mathcal{FV} -like numerical fluxes. This, as remarked in [8, 10, 9], has the effect of spoiling their residual character. Moreover, the design of non-oscillatory \mathcal{DG} schemes relies either on the use of \mathcal{FV} limiters, which can reduce dramatically their accuracy or, as stabilized \mathcal{FE} schemes, on the use of discontinuity capturing operators [97, 166, 72, 19]. This technique basically reduces to adding strongly dissipative terms in localized regions where the gradient of the solution is large. This approach, if on one hand allows to prove the global L^∞ stability of the solution [166], on the other hand *does not* fully guarantee its local monotonicity. More importantly, these shock-capturing (\mathcal{SC}) terms depend on tunable constants which are difficult to determine in a general way. The \mathcal{RD} method, while based on a variable representation similar to the one used in the standard \mathcal{FE} approach, allows to design nonlinear schemes with a true residual character and at the same time guaranteeing *by construction* the preservation of the local monotonicity of the approximation.

As a motivational example, we compare three different schemes on the solution of a 2D conservation law which is a variant of Burger's equation:

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

where the sub-scripts denote partial derivatives. We consider the solution of the last equation on a spatial domain given by a square of side 2, with the initial state depicted on the left on figure 1.1: at time $t = 0$, u is zero everywhere, except in a small square of side 0.5, in which $u = 1$. The final time of the simulation is $t = 1$. On the right, on figure 1.1, we report a reference solution computed with the nonlinear \mathcal{RD} scheme proposed in [8] on a fine unstructured mesh ($\Delta x \approx 1/160$). The initial discontinuity evolves in a twodimensional solution, composed of an *expansion-like* structure (straight lines), across which the solution is piecewise linear, and of a curved discontinuity propagating toward the upper-right corner of the domain (*thick curve*).

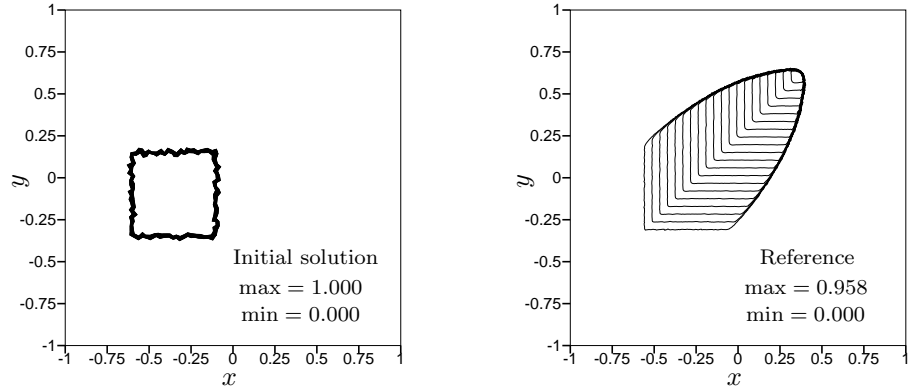


Figure 1.1: 2D Burger's equation. Initial (left) and reference (right) solution at $t = 1$

On figure 1.2 we report contour plots of numerical results obtained, on a coarser unstructured mesh ($\Delta x \approx 1/40$), with the \mathcal{RD} scheme of [8], with a stabilized \mathcal{FE} scheme (Taylor-Galerkin + SUPG + shock capturing (\mathcal{SC})), and with the \mathcal{FV} scheme of [24] (with second order Runge-Kutta time integration).

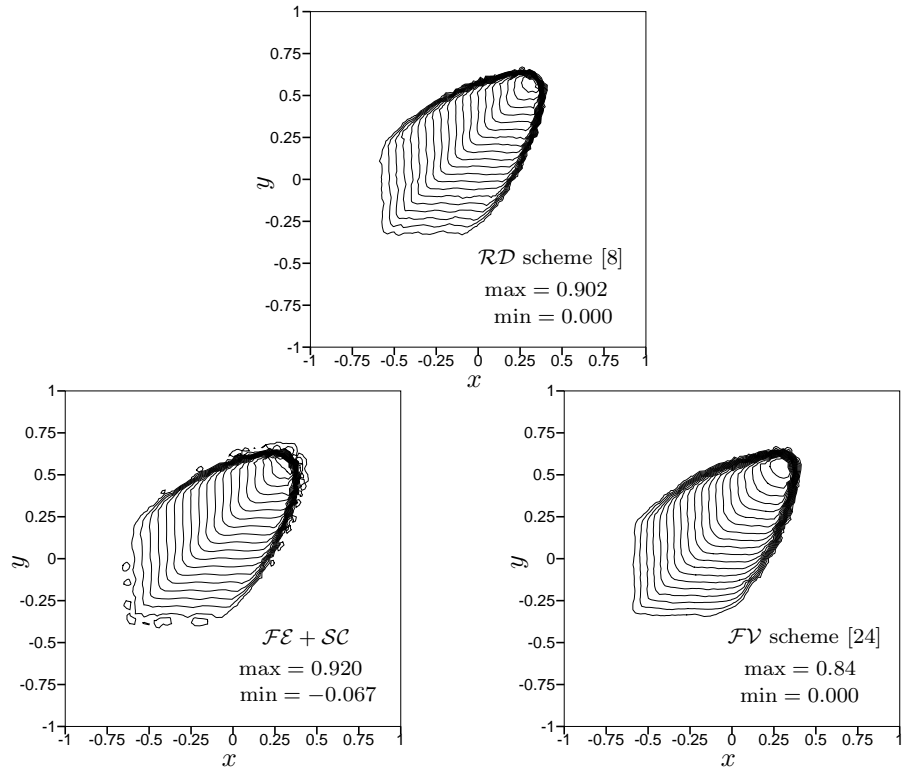


Figure 1.2: 2D Burger's equation: isolines of numerical solutions at $t = 1$. Top: \mathcal{RD} scheme [8]. Bottom-Left: \mathcal{FE} scheme with \mathcal{SC} operator. Bottom-Right: \mathcal{FV} scheme [24]

The global features of the solution are captured by all methods. However, the \mathcal{FV} scheme gives a poorer resolution of both the expansion (the lines are not as straight as in the other plots) and of the discontinuity, which is thicker than in the other results. The \mathcal{FE} results are better, however negative values and small oscillations are obtained. The \mathcal{RD} solution instead shows a very sharp and monotone capturing of the discontinuity. The expansion lines remain straight for longer than in the \mathcal{FE} solution, almost until the line of symmetry of the solution. The analysis is confirmed by the plots on figure 1.3, where the numerical solutions on the line $y = 0.4$ are compared with the reference. We see the poorer resolution of the \mathcal{FV} scheme and the under-shoots in the \mathcal{FE} solution. It might be argued that the finite element results could be improved by a more careful choice of the \mathcal{SC} parameter, governing the amount of *nonlinear* dissipation introduced across the discontinuity. However, this is precisely the reason why the \mathcal{RD} scheme of [8] is more convenient. It guarantees the preservation of the monotonicity of the solution, while being as compact and accurate as the \mathcal{FE} scheme, and more accurate than the \mathcal{FV} scheme. Most importantly, it is completely parameter free.

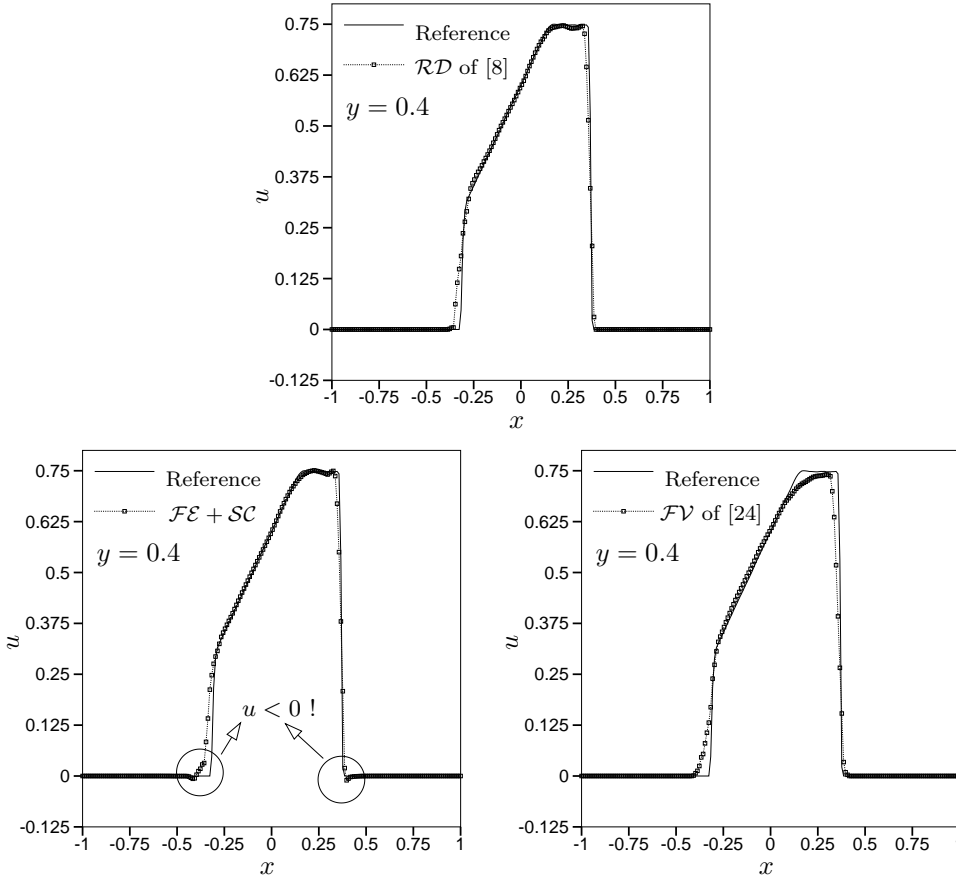


Figure 1.3: 2D Burger's equation: solution at $y = 0.4$ and $t = 1$. Top: \mathcal{RD} scheme [8]. Bottom-Left: \mathcal{FE} scheme with \mathcal{SC} operator. Bottom-Right: \mathcal{FV} scheme [24]

1.2 Objectives and scopes of the present work

The objective of this thesis is to construct and analyze discretization techniques for the numerical solution of conservation laws on unstructured grids. These methods are based on the concept of Residual Distribution (\mathcal{RD}), introduced in [150] by P.L. Roe who refers to it as to the of *Fluctuation-Splitting* (\mathcal{FS}) approach. In this manuscript, we refer equivalently to this methodology as \mathcal{RD} or \mathcal{FS} . The work presented in this thesis unifies results obtained in the last years at the *von Karman Institute for Fluid Dynamics* and at the *Université de Bordeaux I* and, more importantly, proposes new developments leading to accurate and robust solution algorithms for very general steady and time-dependent \mathcal{CL} s. The algorithms are systematically analyzed, tested and, when possible, compared to more traditional schemes. The simple motivational example discussed before shows some of the advantages of the \mathcal{RD} approach. The objective of this thesis is to propose a general conservative formulation of \mathcal{RD} able to handle complex \mathcal{CL} s. At the same time we try to provide an improved understanding of the properties of upwind \mathcal{FS} discretizations. Lastly, the performances of the new conservative discretization procedure are evaluated in a very extensive way, and by keeping a fair and honest eye on the results. Hereafter we recall the background of our work, give further motivation for the study presented and finally discuss in some detail its major contributions.

1.2.1 Historical overview and literature survey on \mathcal{RD}

The fluctuation splitting concept, introduced by P.L. Roe in the early eighties [150], has opened the way to a totally new generation of schemes for the solution of conservation laws on unstructured meshes. Since the very first multidimensional upwind schemes developed by Roe [148, 149], and Roe and Sidilkover [151, 160], \mathcal{FS} schemes have grown thanks to the efforts of the research groups of the University of Michigan [172, 117, 122, 124, 123, 136], led by Roe, of the von Karman Institute for Fluid Dynamics [163, 54, 58, 154, 155, 127, 129, 130, 28, 100, 155, 179, 48, 175, 53, 50, 51, 134, 65], under the supervision of H. Deconinck, of the Université de Bordeaux [4, 7, 119, 3, 8, 10, 6, 9], under the guidance of R. Abgrall, of the *Politecnico di Bari* [62, 61, 63, 60, 41, 40], under the lead of M. Napolitano, of the University of Leeds (M. Hubbard) [91, 92, 76] and of Lund (D. Caraeni) [33, 37], and of many others [87, 86, 186], often in a collaborative effort [131, 39, 165, 164, 139, 56, 57, 120, 12, 90]. The schemes have been proved to be accurate and robust enough to represent a real alternative to \mathcal{FV} and \mathcal{FE} schemes for the computation of steady compressible flow on unstructured meshes. Their higher accuracy and compact character makes them very efficient when compared to \mathcal{FV} schemes [186, 49], especially on parallel platforms [172, 99, 176]. Moreover, the possibility of constructing parameter free non-oscillatory schemes leads to an increased reliability, with respect to \mathcal{FE} . Preliminary results on more complex flow models such as the Magneto-Hydrodynamics (MHD) equations, or two-phase flow models have confirmed this potential.

1.2.2 Open issues

Some important issues still need to be studied, in order for \mathcal{RD} schemes to be competitive with the \mathcal{DG} and RBC schemes appeared in literature lately, and to be able to correctly approximate weak solutions of steady and time-dependent systems of practical interest such as the Euler equations for gases in thermochemical equilibrium or with more complex forms of thermodynamics, or multi-phase flow models. Among these issues, the most relevant ones are an efficient, stable and accurate extension of the method to the time-dependent case, a conservative formulation allowing to handle general forms of thermodynamics, the development of a well-understood procedure for the systematic construction of nonlinear high-order schemes yielding a stable and non-oscillatory approximation of discontinuities, a consistent procedure to extend these nonlinear schemes to inhomogeneous problems, viscous problems and to accuracy higher than two. A discussion of these issues is given in the following subsections.

Residual distribution for time-dependent problems

The use of the \mathcal{FS} approach for time-dependent simulations has seen a very strong progress in the last years and it is still an intense research topic. The main objective of this research is the construction of a framework within which it is possible to obtain discretizations retaining the residual character of steady \mathcal{RD} , as well as to design linear first-order schemes with stable and non-oscillatory shock capturing properties, to be used as a basis for the construction of monotone nonlinear high-order discretizations. It has been always known that in their basic formulation \mathcal{RD} schemes cannot be more than first order accurate in time-dependent computations, due to an inconsistent spatial discretization. Early attempts to cure this problem have resorted to a *Petrov-Galerkin* (\mathcal{PG}) \mathcal{FE} formulation leading to the introduction of a \mathcal{FE} mass-matrix [113, 71]. This approach is very effective in the construction of linear second-order schemes but it leaves open the issue of the construction of non-oscillatory discretizations, since the \mathcal{PG} analogy does not apply to linear monotone \mathcal{RD} schemes. Similarly, in [34, 35, 36, 33, 37, 38], Caraeni and his collaborators have presented schemes in which the time-derivative is consistently included in the definition of the residual. The authors are able to construct in this way second-order schemes for time-dependent calculations and have also proposed an extension of their approach allowing to achieve third-order of accuracy on structured meshes. As in the case of the \mathcal{PG} schemes, this technique does not generalize to linear monotone \mathcal{FS} schemes and hence it does not allow to construct non-oscillatory approximations of discontinuous solutions. Both the \mathcal{PG} schemes and the schemes of Caraeni can be shown to belong to a general family of discretizations making use of a \mathcal{FE} -like mass-matrix consistent with the spatial discretization [62, 61]. Two more approaches can be found in literature to apply \mathcal{RD} schemes for unsteady simulations. The first relies on the use of the \mathcal{FS} formulation of the Lax-Wendroff (LW) scheme [129, 90, 60, 143]. This scheme can be shown to be second order in space and time on structured meshes [143]. However, this property cannot be proved on truly unstructured meshes on which, as we will show, a consistent construction leads also for the LW scheme to the introduction of a mass-matrix.

The study of \mathcal{FS} -type discretizations of time-dependent \mathcal{CL} s has led at last to the *space-time* formulation of \mathcal{RD} . Two different, though similar in spirit, research lines have appeared in literature. One is due to the work reported in [46, 53, 47] in which the authors have written the solution of the time-dependent problem as a sequence of steady problems on space-time *slabs* discretized with space-time linear elements (triangles for a 1D \mathcal{CL} and tetrahedra for 2D problems). The use of *standard* \mathcal{RD} schemes in each space-time slab, allows to obtain discretizations retaining *all* the properties of the steady schemes. The *time-marching* character of the procedure is guaranteed by the use of upwind \mathcal{RD} schemes and by the satisfaction of a time-step constraint. In the references, the authors also propose a *double-layer* formulation in which, by solving at once for the values of the unknown in two successive time levels, *unconditionally* monotone and second order nonlinear schemes can be designed. Although this approach allows to construct schemes with all the desired properties and to make use of all the *numerical artillery* developed for the steady case, it has the drawback of being inherently complex and expensive due to the introduction of time as an additional *independent* unknown, to the generation and storage of the space-time mesh, and to the need of solving for a number of unknowns larger than the number of nodes in the spatial grid, even in its *single-layer* formulation. A space-time formulation of residual distribution making use of *prismatic* space-time elements has been instead proposed in [7, 119, 8]. In the references, the authors design both linear second-order and linear non-oscillatory \mathcal{FS} schemes for time-dependent \mathcal{CL} s. As before, the space-time formulation allows to make use of all the tools developed for steady simulations. The use of prismatic elements, guarantees that the number of unknowns is, in the basic formulation of the method, equal to the number of nodes of the grid. At steady-state, the schemes of [7, 119, 8] reduce to known steady \mathcal{RD} schemes. However, the linear first-order monotone schemes they propose are constrained by a time-step limitation. Using a double layer formulation similar to the one introduced in [46, 53, 47], this limitation has been overcome in [119, 8, 120], where unconditionally monotone and stable nonlinear second-order schemes are presented. Note that the extension of the space-time schemes of [46, 53, 47] to prismatic space-time meshes has been reported in [51]. The framework proposed in [119, 8, 120] and [51] is, at the moment, the only one allowing to design schemes retaining all the properties of steady \mathcal{RD} . It also allows to benefit from the tools developed for steady calculations and to construct linear and nonlinear non-oscillatory schemes in a natural and consistent way.

Residual distribution and conservation

The need of a generalized conservative formulation of \mathcal{RD} stems from the fact that the computation of discontinuous solutions free of numerical oscillations heavily relies on the use of the first-order N scheme [177, 148, 149]. This scheme makes extensive use of the *non-conservative* quasi-linear form of the equations and cannot be conservative unless a multidimensional Roe linearization is used for the flux Jacobians. Unfortunately, this linearization is only available on simplicial elements and when the underlying thermodynamics are simple [56]. This has limited the application of \mathcal{RD} mainly to the computation of flows of perfect gases on triangular (in 2D) and tetrahedral (in 3D) meshes. The first attempts to solve this issue have been based on *ad-hoc* corrections of

a non-conservative formulation of the scheme, as for example in [92, 48]. However, the way in which these corrections have to be included into the discretization is somewhat arbitrary. A more consistent framework has been proposed in [4]. In the reference, the authors introduce a class of non-conservative discretizations based on Gauss volume integration of the quasi-linear form in entropy variables. Using the properties of the Gaussian integration they are able to prove that their schemes indeed converge to the correct weak-solutions. They show how to use their approach to design an N scheme based on the adaptive quadrature of the quasi-linear form of the Euler equations in symmetrizing variables. This technique can be extended to any system of conservation laws with a convex entropy extension, thus solving the problem of the application of \mathcal{RD} in absence of a conservative linearization. The approach of [4] is based on sound mathematical arguments. Global entropy stability on fine meshes can be shown for the N scheme proposed in the reference, while its L_∞ stability can be proved using a wave decomposition technique [10, 9, 118]. The numerical results confirm the theoretical analysis performed in the paper. However, this technique has the drawback of being quite expensive since the number of quadrature points needed to achieve the correct approximation of a shock can be large. A simpler, yet effective, technique has been proposed in [50]. The idea is to approximate directly the integral form of the equations to define the residual. In this way discrete conservation is always guaranteed, provided that a *consistency* constraint is respected. The authors have introduced the terminology \mathcal{CRD} to denote their schemes, indicating that conservation is guaranteed by the definition of the residual as the contour integral of the fluxes on the boundary of the elements of the grid, as opposed to the \mathcal{LRD} schemes for which conservation is guaranteed by the conservative linearization. In the paper it is shown how to construct a conservative variant of the N scheme which does not need a Roe linearization. When applied to the Euler equations, this \mathcal{CRD} N scheme shows performances identical to the standard \mathcal{LRD} N scheme of [177]. Compared to the scheme of [4], the \mathcal{CRD} N scheme is more efficient and computationally cheaper due to the fact that a few Gaussian points are needed on each edge of the grid elements to compute the residual, while the flux Jacobians are evaluated in a single state. No particular theoretical properties have been proved for the \mathcal{CRD} N scheme. The application of this technique to the solution of the ideal MHD equations is shown in [50], while its use to construct \mathcal{FS} schemes on meshes composed of quadrilateral elements has been reported in [134, 63].

Design of high-order nonlinear \mathcal{RD} schemes

Nonlinear schemes are needed to combine high-order of accuracy and monotonicity, as stated by Godunov's theorem [77]. Unfortunately, the construction of high-order nonlinear \mathcal{FS} schemes for systems is yet another open problem. The success of the PSI scheme of Struijs [162] for the solution of scalar steady advection is still far from being achieved for nonlinear systems and issues of robustness and generality are still to be solved. Different techniques can be found in literature. One of these is based on the combination of a linear second-order scheme with a linear monotone scheme (usually the N scheme) through some variant of the *Flux-Corrected-Transport* (FCT) technique [71, 62, 61, 90, 60, 143, 190]. The main problem with this approach is that it generally shows a lack of robustness and it is theoretically very unsatisfactory due to its non-

compact character. A different way of constructing nonlinear \mathcal{RD} discretizations is to blend locally the N scheme with a second-order linear scheme. The local nature of the blending preserves the compactness, while a proper design of the blending function can guarantee both high-order of accuracy and a non-oscillatory approximation of shocks. The definition of this function is not easy. An *ad-hoc* definition, which however has proved to be numerically very effective, is proposed in [177, 154], while a more involved construction, based on positivity and entropy stability considerations, can be found in [3]. In terms of accuracy these *blended* schemes are very competitive with high-order finite volume schemes [3, 9]. Nevertheless, a more robust and general approach has been proposed lately by Abgrall *et al.* in [10, 8, 9, 118, 12]. The basic idea of the technique introduced in the references is to generate nonlinear schemes by locally mapping the residual of a linear non-oscillatory scheme. The nonlinearity is hidden in the mapping which has the property of preserving the sign of its arguments. In simple cases, this can be shown to preserve the monotone character of the discretization. The application of this technique to steady and time-dependent conservation laws has shown improved robustness and accuracy with respect to the blending approach [10, 9, 118, 8].

Extension to inhomogeneous and viscous problems

Real life applications involve the solution of systems of \mathcal{CL} s containing (physical) viscous dissipation and source terms. In order to be able to successfully extend \mathcal{FS} schemes to the solution of inhomogeneous and viscous conservation laws, a proper theoretical framework needs to be developed. In the case of homogeneous problems, the mathematical definition of monotonicity and of a monotone scheme is not easy, even for simple (scalar) \mathcal{CL} s. This is due to the different structure of the mathematical problem that, strictly speaking, does not express anymore simple conservation of the unknown, due to the interaction with the forcing term. As a consequence the design of non-oscillatory discretizations is not at all trivial. This topic has been rarely considered in the \mathcal{RD} literature. Some comments can be found in [151, 58], but a systematic study has never been attempted.

Concerning the extension to viscous problems, the analogy between \mathcal{RD} schemes and \mathcal{FE} has allowed in the past to devise schemes for advective-diffusive problems and, ultimately, for the Navier-Stokes equations. However, as shown by the work of [124], the interaction between the discrete transport (advective) operator and the discrete viscous operator has not been properly addressed in the major \mathcal{FS} literature. The extension to the time-dependent case has never been studied, in particular for the nonlinear space-time \mathcal{RD} schemes recently developed [8, 51].

Very high-order \mathcal{RD} schemes

The \mathcal{FS} approach gives a natural framework for the construction of schemes of accuracy higher than two, as shown by the results of [12, 9, 139]. However, the full capabilities of the schemes are yet to be exploited. In [12, 9] third and fourth order schemes

are presented, together with a construction allowing to achieve *any* order of accuracy. Unfortunately, the nonlinear schemes proposed in the references are far from retaining all the properties of their second-order versions, and work is still needed in this sense.

1.2.3 The contribution of this thesis

This thesis is an attempt to deal with some of the issues described in the previous subsections by combining and analyzing some of the ideas present in literature, at the same time proposing new concepts. As an output of this process, we propose a very general conservative framework within which we construct non-oscillatory high-order nonlinear \mathcal{RD} schemes for steady and time-dependent conservation laws. Our approach is based on three main elements:

1. A positivity theory allowing to design discretizations verifying a discrete maximum principle and uniform L^∞ stability;
2. A technique allowing to generate, starting from first-order linear positive schemes, nonlinear positive second-order schemes;
3. A conservative formulation enabling to extend these schemes to nonlinear \mathcal{CL} s.

The development of each of these design stages has led to the following main contributions of our work.

- Introduction of a general framework for the positivity and energy stability analysis of explicit and implicit compact cell-vertex schemes for the solution of the advection equation on unstructured meshes. The positivity analysis includes solution independent source terms;
- Detailed analysis of \mathcal{FS} discretizations, including a discussion on geometrical constructions allowing to design positive nonlinear schemes. Several yet unpublished properties of these schemes are discussed;
- Construction of nonlinear positive \mathcal{RD} schemes by means of the mapping technique proposed in [10, 12]. This approach allows to construct, in more general settings, analogs of the PSI scheme of Stuijs [165]. In this thesis we analyze the stability and the well-posedness of these schemes;
- The extension to scalar nonlinear \mathcal{CL} s, is achieved by using the \mathcal{CRD} technique [50]. The positivity and the (entropy) dissipation properties of the schemes are studied;
- Extension of the conservative schemes to the approximation of time-dependent weak solutions of \mathcal{CL} s;

- Construction of nonlinear, high-order, non-oscillatory, conservative schemes for the solution of steady and time-dependent \mathcal{CL} s on unstructured meshes. This is achieved by applying the limiting approach of [10, 12] to conservative \mathcal{CRD} variants of the linear N scheme. The conservative formulation of the N scheme guarantees the well-posedness of the procedure;
- Evaluation of the schemes on the Euler equations;
- Evaluation of the schemes on a two-phase flow model;
- Evaluation of the schemes on the shallow-water equations. The \mathcal{RD} approach developed here is proved to *preserve exactly* the so-called lake-at-rest solutions;

This work means to asses in an *honest way* the performances of the schemes proposed, as well as of the *available \mathcal{RD} technology*, employed to construct them. By no means the thesis aims at convincing the reader that what we propose is the absolute best at the moment. Our objective is to point out clearly the main advantages of the \mathcal{RD} approach while trying to understand, by analytical means and numerical simulations, the main weaknesses of the constructions presented. The thesis focuses on the case of twodimensional \mathcal{CL} s, however, the theory easily generalizes to three space dimensions.

1.3 Structure of the manuscript

The organization of the manuscript has been conceived keeping in mind the modeling steps which lead, starting from a physical problem, to a discrete solution verifying certain properties. In particular, the idea behind the structure of the thesis is to first present the continuous problem which needs to be solved, then to introduce the framework of a discrete space and discrete unknowns and finally to discuss and validate the discretization approach. It is hoped that this translates in a unique flow of concepts allowing to understand why some analytical tools are used and on what grounds some properties are claimed to be important. To this end we have chosen to

1. Introduce in **Chapter 2** the abstract model of a hyperbolic conservation law. At this stage, the most general case considered is that of hyperbolic problem with a *solution independent* source term. This chapter serves to present the definition of the *weak* exact solution to the continuous problem and to show that this solution is subject to certain *stability* constraints, which can guarantee its existence¹, uniqueness and physical relevance. This chapter also serves to introduce part of the notation used in the thesis;
2. Present in **Chapter 3** the preliminary steps needed to translate the continuous problem into an algebraic equation: mesh and discrete unknowns. Also this chapter serves to introduce part of the notation used throughout the manuscript;

¹in the cases in which it can be proved

3. Introduce in **Chapter 4** an abstract prototype for compact cell-vertex discretizations of steady linear scalar problems. The formalism used in this chapter is tailored for \mathcal{RD} schemes, however it encompasses also other schemes of the \mathcal{FE} and \mathcal{FV} type. The goal of this chapter is to show under which conditions the discrete solution enjoys stability properties similar to the ones of the exact solution. The content of this chapter is somewhat abstract, it allows however to shorten the analysis of the schemes presented in the next chapters and to introduce, in the linear scalar case, concepts which can then be formally extended to the more complex cases considered later;
4. Describe in **Chapter 5** the basics of the \mathcal{FS} approach. This part of the thesis is meant to introduce the basic philosophy behind the construction of \mathcal{FS} discretizations. We focus on linear scalar advection, trying to pay as much as possible attention to geometrical aspects of the discretization, so to give a more understandable presentation. However, theoretical issues are covered as well. This chapter also allows to introduce in a simpler way the construction of nonlinear \mathcal{RD} . Illustrative computational examples are given;
5. Discuss in **Chapter 6** the extension of the schemes to the solution of *scalar* steady nonlinear \mathcal{CL} s. This chapter allows to present conservative \mathcal{RD} discretizations and analyze their properties in the relatively simple framework of a scalar problem, simplifying the task of their analysis in the system case, where many concepts extend formally. Illustrative computational examples are given;
6. Discuss in **Chapter 7** the consistent extension of the conservative schemes to time-dependent nonlinear scalar conservation laws. As the previous one, this chapter allows to discuss and analyze the schemes in a relatively simple framework. Illustrative computational examples are given;
7. Describe in **Chapter 8** the extension of the schemes to systems. The extension is presented first for linear symmetric systems and then for systems of nonlinear \mathcal{CL} s. Few additional theoretical results have to be introduced in this chapter, while other properties can be extended formally or with little modifications;
8. Discuss in **Chapter 9** some issues related to the implementation of the schemes in the case of systems. This chapter also allows to summarize the whole discretization strategy and clearly describe the schemes actually used for the applications of the following chapters;
9. Evaluate the schemes on the solution of the Euler equations for a perfect gas in **Chapter 10**. Several test problems are considered and, when possible, comparisons are made with \mathcal{FV} schemes;
10. Evaluate the schemes on the solution of a two-phase flow model in **Chapter 11**. Several test problems are considered;
11. Evaluate the schemes on the solution of the shallow-water equations in **Chapter 12**. In this case, the schemes developed in this thesis can be shown to have some very important properties, related to the type of solutions the shallow-water equations admit. After discussing these theoretical aspects, several test problems are considered;

12. Summarize the results of the thesis and its main achievements. Underline the current limitations of the approach proposed and show some of the ways left open by this work together with the ones opened, proposing some possible routes to improve and extend the work presented.

Almost every chapter is ended by a summary recalling the main results and ideas presented. It is evident that the thesis has not an *old-vs-new* structure, new and old results being interlaced in the manuscript in order to obtain a coherent presentation. Practically all the analysis is performed in the scalar case, which is the only one in which a rigorous theoretical framework can be developed, the analysis of systems being more technical and less well understood. This probably makes the first half of the manuscript somewhat hard to digest. However, it allows to present the theory in a simpler setting, where geometrical analogies often help its understanding. It also allows to split this thesis in two halves which can be read almost independently:

- (a) The reader interested in the theoretical analysis and construction of the schemes will find all the details in **Chapter 2** to **Chapter 8**;
- (b) The reader interested in the numerical applications can skip the first eight chapters. **Chapter 9** gives an in depth summary of the computational techniques used allowing to go directly to **Chapter 10**, **Chapter 11** and **Chapter 12** devoted the evaluation of the schemes on different systems of \mathcal{CL} s. If needed, references to the appropriate sections in the first half of the thesis are given.

Chapter 2

Conservation laws: continuous problem and related stability

This thesis consider the numerical approximation of solutions of the following system of conservation laws:

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F} = \mathcal{S}(x, y) \quad \text{on} \quad \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^d \times \mathbb{R}^+, \quad (2.1)$$

where $\mathbf{u} : \Omega_T \mapsto \mathbb{R}^m$ is an m -vector of conserved quantities, $\mathcal{F} : \mathbb{R}^m \mapsto \mathbb{R}^{m \times d}$ is the tensor of the conservative fluxes, $\mathcal{S} : \Omega \mapsto \mathbb{R}^m$ is an m -vector of source terms independent on \mathbf{u} and $\Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^d \times \mathbb{R}^+$ is the space-time domain over which solutions are sought. System (2.1) is also equipped with a set of *boundary conditions* (BC) on $\partial\Omega_T$ (or on properly defined portions of this set), and with an initial solution

$$\mathbf{u}(x_1, \dots, x_d, t = 0) = \mathbf{u}_0(x_1, \dots, x_d). \quad (2.2)$$

Throughout the thesis, we focus on the twodimensional case $d = 2$ and

$$\mathcal{F}(\mathbf{u}) = (\mathbf{F}(\mathbf{u}), \mathbf{G}(\mathbf{u})), \quad \text{and} \quad \vec{x} = (x_1, x_2) = (x, y), \quad (2.3)$$

however the theory presented easily extends to three space dimensions. We assume that the system is *hyperbolic*, that is, for any given direction $\vec{\xi} = (\xi_1, \xi_2) \in \mathbb{R}^2$, the matrix

$$K(\vec{\xi}, \mathbf{u}) = \frac{\partial \mathbf{F}(\mathbf{u})}{\partial \mathbf{u}} \xi_1 + \frac{\partial \mathbf{G}(\mathbf{u})}{\partial \mathbf{u}} \xi_2 \quad (2.4)$$

admits a complete set of real eigenvalues and linearly independent eigenvectors. As we will see shortly, this condition can be deduced by a more general property of the system, however it has been postulated here for sake of clarity. In particular, everywhere in the text, we shall denote by $\Lambda(\vec{\xi}, \mathbf{u})$ the diagonal matrix of the eigenvalues of $K(\vec{\xi}, \mathbf{u})$ and by $R(\vec{\xi}, \mathbf{u})$ the matrix of its right eigenvectors, so that

$$K(\vec{\xi}, \mathbf{u}) = R(\vec{\xi}, \mathbf{u}) \Lambda(\vec{\xi}, \mathbf{u}) R(\vec{\xi}, \mathbf{u})^{-1}.$$

We also introduce the matrices

$$\Lambda(\vec{\xi}, \mathbf{u})^+ = \text{diag} \{ \max(0, \lambda_k) \}_{k=1}^m \quad \Lambda(\vec{\xi}, \mathbf{u})^- = \text{diag} \{ \min(0, \lambda_k) \}_{k=1}^m$$

and

$$|\Lambda(\vec{\xi}, \mathbf{u})| = \text{diag} \{ |\lambda_k| \}_{k=1}^m = \Lambda(\vec{\xi}, \mathbf{u})^+ - \Lambda(\vec{\xi}, \mathbf{u})^- ,$$

where λ_k is the k -th eigenvalue of $K(\vec{\xi}, \mathbf{u})$. The positive and negative parts and the absolute value of $K(\vec{\xi}, \mathbf{u})$ are then defined as

$$K(\vec{\xi}, \mathbf{u})^\pm = R(\vec{\xi}, \mathbf{u}) \Lambda(\vec{\xi}, \mathbf{u})^\pm R(\vec{\xi}, \mathbf{u})^{-1}, \quad |K(\vec{\xi}, \mathbf{u})| = R(\vec{\xi}, \mathbf{u}) |\Lambda(\vec{\xi}, \mathbf{u})| R(\vec{\xi}, \mathbf{u})^{-1}. \quad (2.5)$$

Clearly one has

$$K(\vec{\xi}, \mathbf{u}) = K(\vec{\xi}, \mathbf{u})^+ + K(\vec{\xi}, \mathbf{u})^-, \quad |K(\vec{\xi}, \mathbf{u})| = K(\vec{\xi}, \mathbf{u})^+ - K(\vec{\xi}, \mathbf{u})^- . \quad (2.6)$$

The objective of this chapter is to characterize solutions of (2.1). We will briefly recall some *stability* conditions allowing, in some cases, to guarantee the existence of such solutions and their *physical relevance*. This, is believed, will be useful to justify and understand the need of requiring certain properties to be satisfied by the discrete solution. Since the thesis also considers simpler \mathcal{CL} s, the discussion will be done in an increasingly level of complexity, starting from the simplest first-order scalar transport equation to arrive at the end at (2.1). Throughout the text we try to focus on the aspects which are in practice most relevant for the analysis of the schemes. The discussion is far from being a general review of the theory of the solutions of conservation laws for which an extensive bibliography is given in the text. In particular, several aspects related to functional analysis are omitted or simplified giving however appropriate references where a rigorous presentation can be found. As done for (2.1), the notation used throughout the manuscript is introduced along the discussion.

2.1 The scalar advection equation

We start considering the scalar advection equation

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \mathcal{S}(x, y) \quad \text{on} \quad \Omega_T \subset \mathbb{R}^2 \times \mathbb{R}^+, \quad (2.7)$$

with $\vec{a} = (a_1, a_2) \in \mathbb{R}^2$ constant. Equation (2.7) can be recast in a form similar to (2.1), assuming $m = 1$, $\mathbf{u} = u \in \mathbb{R}$, $\mathcal{F}(\mathbf{u}) = \mathcal{F}(u) = \vec{a} u$ and $\mathcal{S}(x, y) = \mathcal{S}(x, y)$. Given a smooth initial solution $u_0(x, y)$, a regular enough function¹ $u(x, y, t)$ verifying (2.7) in a pointwise manner and such that $u(x, y, 0) = u_0(x, y)$, is called a *classical solution*. Exact solutions to this problem can be precisely represented [70, 174, 67, 14, 137]. In the homogeneous case $\mathcal{S}(x, y) = 0$, one easily shows that (2.7) is equivalent to

$$\frac{du(x(t), y(t), t)}{dt} = 0$$

¹ $u_0(x, y) \in C^1(\mathbb{R}^2)$ and $u(x, y, t) \in C^1(\Omega_T)$ is a natural requirement for this problem

on the so-called *characteristic curves* $\Gamma_{\vec{\zeta}}$ parametrized by $(x(t), y(t), t)$ with

$$\frac{d(x(t), y(t))}{dt} = \vec{a}, \quad (x(0), y(0)) = \vec{\zeta}. \quad (2.8)$$

The solution is constant along characteristic curves (in this case straight lines) and can be written in closed form as [70, 174, 67, 14, 137]

$$u(x(t), y(t), t) = u_0(\vec{\zeta}). \quad (2.9)$$

This corresponds to the propagation of the initial data in space-time along the direction $(\vec{a}, 1) \in \mathbb{R}^2 \times \mathbb{R}$. These solutions can be further characterized as follows:

Maximum principle Since the initial data are simply propagating in space-time, one has trivially

$$\inf_{\Omega} u_0(x, y) \leq u(x, y, t) \leq \sup_{\Omega} u_0(x, y). \quad (2.10)$$

Last inequality is known as the *maximum principle*.

Energy conservation (and stability) Simple arguments (see *e.g* [152]) can be used to show that¹ the following principle of *conservation of energy* holds for the solution:

$$\|u(t)\|_{L^2} = \|u_0\|_{L^2}, \quad (2.11)$$

where $\|(\cdot)\|_{L^2}$ denotes the standard L^2 norm on Ω :

$$\|(\cdot)\|_{L^2}^2 = \int_{\Omega} (\cdot)^2 dx dy.$$

Note that energy growth in time would correspond to an unstable behavior, while *energy stability* implies the inequality [70, 174, 152, 67, 14, 137]

$$\|u(t)\|_{L^2} \leq \|u_0\|_{L^2}, \quad (2.12)$$

The stability associated to (2.12) corresponds to the presence of a *dissipative* phenomenon [70, 174, 152, 67, 14, 137].

Inhomogeneous case Consider now the case $\mathcal{S} = \mathcal{S}(x, y) \neq 0$, with

$$\sup_{\mathbb{R}^2} |\mathcal{S}(x, y)| < \infty.$$

It is easy to check that in this case the exact solution becomes

$$u(x, y, t) = u_0(\vec{\zeta}) + \int_0^t \mathcal{S}(x(s), y(s)) ds,$$

with $(x(s), y(s))$ respecting (2.8). In this case, we do not have, strictly speaking, a maximum principle, however at $t_f < \infty$, the following stability estimate holds

$$\inf_{\Omega} u_0(x, y) + t_f \inf_{\Omega} \mathcal{S}(x, y) \leq u(x, y, t_f) \leq \sup_{\Omega} u_0(x, y) + t_f \sup_{\Omega} \mathcal{S}(x, y) \quad (2.13)$$

¹without taking into account the BCs, or assuming homogeneous BCs

Similarly, an energy estimate can be derived (see *e.g.* the lecture notes [152]):

$$\|u(t_f)\|_{L^2(\Omega)}^2 \leq e^{t_f} \|u_0\|_{L^2(\Omega)}^2 + \int_0^{t_f} e^{t_f-s} \|\mathcal{S}(x(s), y(s))\|_{L^2(\Omega)}^2 ds. \quad (2.14)$$

Weaker regularity Even though classical solutions must have enough regularity for (2.7) to hold in each point, it is easily shown that (2.9) makes sense also if the initial data, hence the solution, have little regularity or even discontinuous. Similarly, the bounds on the solution and the energy estimates can be derived with much weaker assumptions. The definition of a *weak solution*, will be given shortly, in the more general context of a nonlinear conservation law.

Note that the theory extends easily to the case $\vec{a} = \vec{a}(x, y) \in C^1(\mathbb{R}^2)^2$ [152].

2.2 Scalar nonlinear conservation laws

Consider now the nonlinear scalar problem

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathcal{F}(u) = 0 \quad \text{on } \Omega_T \subset \mathbb{R}^2 \times \mathbb{R}^+, \quad (2.15)$$

obtained from (2.1) when $m = 1$, $\mathbf{u} = u \in \mathbb{R}$, $\mathcal{F}(\mathbf{u}) = \mathcal{F}(u) = (F(u), G(u)) \in \mathbb{R}^2$ and $\mathcal{S} = \mathcal{S} = 0$. Last equation can be written in the following quasi-linear form

$$\frac{\partial u}{\partial t} + \vec{a}(u) \cdot \nabla u = 0,$$

having introduced the Jacobian vector

$$\vec{a}(u) = \left(\frac{dF(u)}{du}, \frac{dG(u)}{du} \right). \quad (2.16)$$

Last equation is formally similar to (2.7). Classical solutions of (2.15) can be represented exactly as the solutions of the advection equation. In particular, introducing the characteristic curves $\Gamma_{\vec{\zeta}}$ parametrized by $(x(t), y(t), t)$ with

$$\frac{d(x(t), y(t))}{dt} = \vec{a}(u(x(t), y(t), t)), \quad (x(0), y(0)) = \vec{\zeta},$$

a solution of (2.15) is

$$u(x(t), y(t), t) = u_0(\vec{\zeta}).$$

Moreover, since u is constant along each characteristic, so is $\vec{a}(u)$, hence the characteristic curves are straight lines. However, since the slope of the characteristics depends on the initial data $u_0(x, y)$, even for a smooth initial solution, characteristic lines can cross. At the intersections no unique solution can exist in general, hence classical solutions exist up to the time t^* in which two characteristics cross for the first time. This limitation is overcome by introducing a more general concept of solution. The discussion of the next sections, follows [156, 25, 67, 68].

2.2.1 Weak solutions

The notion of a *weak solution* is introduced to cope with the fact that classical solutions exist only until a finite time t^* . A function $u \in L^\infty(\Omega_T)$ is a weak solution if it satisfies (2.15) and the initial condition *in the sense of distributions* [70, 174, 13, 31, 67, 137]:

$$\int_{\Omega_T} \left(u \frac{\partial \varphi}{\partial t} + \mathcal{F} \cdot \nabla \varphi \right) dx dy dt + \int_{\Omega} u_0 \varphi(x, y, 0) dx dy = 0 \quad \forall \varphi \in C_0^1(\Omega_T). \quad (2.17)$$

While classical solutions are also weak solutions of the problem, condition (2.17) enlarges the set of possible admissible solutions to the set of bounded functions. However, additional constraints are needed to guarantee the uniqueness of the the solution. As the particular case $\mathcal{F} = \tilde{a}u$, weak solutions of (2.7) are also defined by (2.17).

2.2.2 Conservation

A whole class of weak solutions is given by functions which are piecewise smooth [156, 25]. Smooth regions are separated by discontinuities which can be characterized precisely. For simplicity, consider the case in which there exist Ω_L and Ω_R , subsets of Ω_T such that

$$\overline{\Omega}_L \cup \overline{\Omega}_R = \overline{\Omega}_T, \quad \Omega_L \cap \Omega_R = \emptyset,$$

with $\overline{\Omega}$ the closure of a set Ω ¹. Given a function $u \in C^1(\Omega_k)$, $k = L, R$, then u is a weak solution of (2.15) if it is a classical solution in Ω_L and Ω_R and if on the surface S_{LR} separating the two sets it respects the following *Rankine-Hugoniot jump condition*:

$$n_{LR}^t [u]_{LR} = [\mathcal{F}(u)]_{LR} \cdot \vec{n}_{LR}, \quad (2.18)$$

with $(\vec{n}_{LR}, -n_{LR}^t) \in \mathbb{R}^2 \times \mathbb{R}$ the space-time unit normal to S_{LR} , and $[(\cdot)] = (\cdot)_R - (\cdot)_L$ the jump of a quantity across S_{LR} in the direction of $(\vec{n}_{LR}, -n_{LR}^t)$. Condition (2.18) states the *conservation* of u across a discontinuity and it constrains the speed at which the discontinuity can move for a given jump of the unknown. The jump condition is however not enough to uniquely determine the solution. There are in literature classical examples showing that in some cases there exist infinite sets of piecewise smooth functions respecting (2.18) and (2.15) (see [156, 25, 110, 68] for example).

2.2.3 Entropy and dissipation

Consider the regularized problem

$$\frac{\partial u^\mu}{\partial t} + \nabla \cdot \mathcal{F}(u^\mu) = \mu \Delta u^\mu$$

¹the *smallest* closed set containing Ω (or the intersection of *all* the closed sets containing Ω)

Solutions of (2.15) can be seen as the limit of the solutions of this regularized equation for vanishing values of the *viscosity* $\mu > 0$. The Laplacian operator on the right-hand-side of last equation, models a dissipative phenomenon. For this equation, it is possible to prove that unique smooth solutions do exist [67, 68]. Next, we introduce the concept of a *convex entropy pair* $(\mathcal{H}(u), \mathcal{G}(u))$, where $\mathcal{H}(u)$ is an entropy function and $\mathcal{G}(u)$ the corresponding entropy flux, such that

$$\frac{d\mathcal{G}}{du} = \frac{d\mathcal{H}}{du} \frac{d\mathcal{F}}{du}, \quad \frac{d^2\mathcal{H}}{du^2} > 0 \quad (2.19)$$

Multiplying the regularized problem by $d\mathcal{H}/du$, using (2.19) and integrating by parts one obtains

$$\begin{aligned} \frac{d\mathcal{H}}{du} \frac{\partial u^\mu}{\partial t} + \frac{d\mathcal{H}}{du} \nabla \cdot \mathcal{F}(u^\mu) &= \mu \frac{d\mathcal{H}}{du} \Delta u^\mu \\ \frac{\partial \mathcal{H}(u^\mu)}{\partial t} + \nabla \cdot \mathcal{G}(u^\mu) &= \mu \Delta \mathcal{H}(u^\mu) - \overbrace{\mu \frac{d^2\mathcal{H}}{du^2} \nabla u^\mu \cdot \nabla u^\mu}^{\geq 0} \\ \frac{\partial \mathcal{H}(u^\mu)}{\partial t} + \nabla \cdot \mathcal{G}(u^\mu) &\leq \mu \Delta \mathcal{H}(u^\mu) \end{aligned}$$

If we now take the limit $\mu \rightarrow 0$, the entropy pair respects, in a weak or distributional sense [70, 174, 152, 13, 31, 67, 137]

$$\frac{\partial \mathcal{H}(u)}{\partial t} + \nabla \mathcal{G}(u) \leq 0. \quad (2.20)$$

This leads to the concept of an *entropy weak solution* (or *vanishing viscosity solution*). In particular, a weak solution is said to be an entropy weak solution or vanishing viscosity solution of (2.15) if for all convex entropy pairs $(\mathcal{H}(u), \mathcal{G}(u))$ and $\forall \varphi \in C_0^1(\Omega_T)$

$$\int_{\Omega_T} \left(\mathcal{H}(u) \frac{\partial \varphi}{\partial t} + \mathcal{G}(u) \cdot \nabla \varphi \right) dx dy dt + \int_{\Omega} \mathcal{H}(u_0) \varphi(x, y, 0) dx dy \geq 0. \quad (2.21)$$

It is possible to prove that *entropy solutions are unique* [105, 84, 104, 156, 25, 110].

2.2.4 Max principle

The dissipation mechanism implied by the entropy inequality represents in itself a stability condition which ultimately leads to the uniqueness of weak solutions. Moreover, these solutions can be shown to have a set of properties guaranteeing their boundedness and continuous dependence on the initial data (see [156, 25, 110] and references therein). In particular, they respect the maximum principle (2.10).

2.3 Linear symmetric hyperbolic systems

We now consider a system of PDEs of the form

$$\frac{\partial \mathbf{u}}{\partial t} + A_1 \frac{\partial \mathbf{u}}{\partial x} + A_2 \frac{\partial \mathbf{u}}{\partial y} = \mathcal{S}(x, y) \quad \text{on } \Omega_T \subset \mathbb{R}^2 \times \mathbb{R}^+, \quad (2.22)$$

with constant symmetric matrices $A_j, \forall j$. Note that the hyperbolic character of the system is guaranteed by the symmetry of the matrices. The characterization of the solutions of such a system is more technical than in the case of the advection equation. However, it is known that, in the homogeneous case, the system admits *simple wave* solutions propagating at finite speeds which are related to the eigenvalues of the matrices A_1 and A_2 . We shall just mention that for (2.22), the *existence and uniqueness of the solutions is also proved by resorting to energy estimates* [67, 14, 137, 75, 74]. We recall that for (2.22) energy conservation reads

$$\|\mathbf{u}(t)\|_{L^2(\Omega)}^2 = \|\mathbf{u}_0\|_{L^2(\Omega)}^2, \quad (2.23)$$

while energy dissipation (or stability) implies

$$\|\mathbf{u}(t)\|_{L^2(\Omega)}^2 \leq \|\mathbf{u}_0\|_{L^2(\Omega)}^2. \quad (2.24)$$

Concerning the existence of a maximum principle, as argued in [10, 9], the analysis of the initial value problem given by (2.22) with piecewise smooth initial data [1] shows that the solutions are also piecewise smooth with limited oscillations in correspondence of discontinuities. This justifies the use of analytical techniques aiming at proving that discrete approximations of solutions of (2.22) satisfy some L^∞ stability criterion.

2.4 Nonlinear systems of conservation laws

Finally, we consider the case of a system of conservation laws

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F} = 0 \quad \text{on } \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^d \times \mathbb{R}^+,$$

with $\mathcal{F} = (\mathbf{F}, \mathbf{G})$. We assume that the system is equipped with a convex entropy extension so that one also has the additional *scalar* inequality

$$\frac{\partial \mathcal{H}(\mathbf{u})}{\partial t} + \nabla \cdot \mathcal{G}(\mathbf{u}) \leq 0, \quad (2.25)$$

with $\mathcal{H} : \mathbb{R}^m \mapsto \mathbb{R}$ a *convex* entropy function and $\mathcal{G} : \mathbb{R}^m \mapsto \mathbb{R}^2$ the entropy flux:

$$\frac{\partial \mathcal{G}(\mathbf{u})}{\partial \mathbf{u}} = \frac{\partial \mathcal{H}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathcal{F}(\mathbf{u})}{\partial \mathbf{u}} = \mathbf{v}^T(\mathbf{u}) \frac{\partial \mathcal{F}(\mathbf{u})}{\partial \mathbf{u}}, \quad (2.26)$$

where the superscript T denotes the transpose operator and having introduced the *vector of entropy variables*

$$\mathbf{v}(\mathbf{u}) = \frac{\partial \mathcal{H}(\mathbf{u})}{\partial \mathbf{u}}^T. \quad (2.27)$$

The convexity of $\mathcal{H}(\mathbf{u})$ guarantees that the inverse Hessian matrix

$$A_0 = \left(\frac{\partial^2 \mathcal{H}(\mathbf{u})}{\partial \mathbf{u}^2} \right)^{-1} \quad (2.28)$$

is positive definite and that the mapping $\mathbf{u} \mapsto \mathbf{v}$ is invertible. Note that A_0 is symmetric by definition. The existence of (at least one) convex entropy pair for *physically derived* systems of conservation laws is known [83, 167, 168, 68]. As the scalar conservation law (2.15), a system of \mathcal{CL} s admits smooth classical solutions. In particular, the symmetrization theory for first-order systems of conservation laws (see [78, 121] and also [83, 167, 168] and references therein) ensures that, under the change of variables $\mathbf{u} \mapsto \mathbf{v}$ with \mathbf{v} given by (2.27), the system can be written into the symmetric form

$$A_0 \frac{\partial \mathbf{v}}{\partial t} + A_1 \frac{\partial \mathbf{v}}{\partial x} + A_2 \frac{\partial \mathbf{v}}{\partial y} = 0, \quad A_1 = \frac{\partial \mathbf{F}}{\partial \mathbf{v}}, \quad A_2 = \frac{\partial \mathbf{G}}{\partial \mathbf{v}} \quad (2.29)$$

with A_j symmetric $\forall j = 1, 2$ and A_0 symmetric positive definite and given by (2.28). As in the case of a linear system, the symmetry of the A_j matrices implies the hyperbolic character of the problem. Classical solutions can be characterized as the solution of a linear symmetric system of PDEs. In particular, there exist simple wave-like solutions corresponding to the propagation of the initial data at finite speeds depending on the eigenvalues of the Jacobians A_j . However, as for equation (2.15), the nonlinear character of the system leads in finite time to the formation of singular solutions. To cope with this situation, weak solutions are introduced.

Weak solutions As in the scalar case, a bounded vector function \mathbf{u} with components in $L^\infty(\Omega_T)$ is called a weak solution of the system if $\forall \varphi \in C_0^1(\Omega_T)$

$$\int_{\Omega_T} \left(\mathbf{u} \frac{\partial \varphi}{\partial t} + \mathcal{F} \cdot \nabla \varphi \right) dx dy dt + \int_{\Omega} \mathbf{u}_0 \varphi(x, y, 0) dx dy = 0. \quad (2.30)$$

Conservation A first characterization of weak solutions is obtained by considering piecewise smooth functions. If Ω_L and Ω_R are two subsets of Ω_T such that their intersection is the empty set and that $\Omega_L \cup \Omega_R = \Omega_T$, then \mathbf{u} is a weak solution if $\mathbf{u} \in C^1(\Omega_k)^m$, $k = L, R$, if it is a classical solution in both sub-domains, and if on the surface S_{LR} separating Ω_L and Ω_R it respects (with the notation of (2.18))

$$n_{LR}^t [\mathbf{u}]_{LR} = [\mathcal{F}(\mathbf{u})]_{LR} \cdot \vec{n}_{LR}, \quad (2.31)$$

As before, this condition expresses the conservation of \mathbf{u} across a discontinuity. As easily verified for *e.g.* the Euler equations for a perfect gas, if (2.25) is not taken into account, for a given initial condition several weak solutions can be found. A classical example is that of the so-called *expansion shock* solutions.

Entropy and dissipation Also for a system, the existence of an entropy inequality guarantees that a (stabilizing) dissipative mechanism determines the structure of the solution. To see this, introduce the regularized problem [171]

$$\frac{\partial \mathbf{u}^\mu}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{u}^\mu) = \mu \nabla \cdot (P(\mathbf{u}, \nabla \mathbf{u}) \nabla \mathbf{u}^\mu),$$

with $P(\mathbf{u}, \nabla \mathbf{u})$ a *viscosity matrix* verifying

$$A_0 P = (A_0 P)^T \quad \text{positive definite}$$

with A_0 as in (2.28). Proceeding as in the scalar case, one gets

$$\begin{aligned} \mathbf{v}^T \frac{\partial \mathbf{u}^\mu}{\partial t} + \mathbf{v}^T \nabla \cdot \mathcal{F}(\mathbf{u}^\mu) &= \mu \mathbf{v}^T \nabla \cdot (P A_0 \nabla \mathbf{v}) \\ \frac{\partial \mathcal{H}(\mathbf{u}^\mu)}{\partial t} + \nabla \cdot \mathcal{G}(\mathbf{u}^\mu) &= \mu \nabla \cdot (\mathbf{v}^T P A_0 \nabla \mathbf{v}) - \overbrace{\mu (\nabla \mathbf{v})^T P A_0 \nabla \mathbf{v}}^{\geq 0} \\ \frac{\partial \mathcal{H}(\mathbf{u}^\mu)}{\partial t} + \nabla \cdot \mathcal{G}(\mathbf{u}^\mu) &\leq \mu \nabla \cdot (\mathbf{v}^T P A_0 \nabla \mathbf{v}) \end{aligned}$$

Weak solutions of the nonlinear system of \mathcal{CL} s are then obtained as the limit of the solutions of the regularized problem for vanishing values of the *viscosity* $\mu > 0$. Note that, in this limit, last inequality reduces precisely to (2.25). As in scalar case, the limit is intended in a distributional sense: a weak solution is an entropy weak solution, or vanishing viscosity solution, if for all convex entropy pairs $(\mathcal{H}(\mathbf{u}), \mathcal{G}(\mathbf{u}))$ and $\forall \varphi \in C_0^1(\Omega_T)$

$$\int_{\Omega_T} \left(\mathcal{H}(\mathbf{u}) \frac{\partial \varphi}{\partial t} + \mathcal{G}(\mathbf{u}) \cdot \nabla \varphi \right) dx dy dt + \int_{\Omega} \mathcal{H}(\mathbf{u}_0) \varphi(x, y, 0) dx dy \geq 0. \quad (2.32)$$

In the case of a system of conservation laws, even for entropy weak solutions the question of uniqueness is nontrivial, however see [169, 170, 171].

Maximum principle and inhomogeneous problems For systems of \mathcal{CL} s the question of the existence of a maximum principle is non-trivial. However, the piecewise smooth characterization discussed before allows and justifies, at least in principle, to look for numerical techniques guaranteeing limited (or no) oscillations at discontinuities. The same arguments hold for the inhomogeneous case

2.5 Summary

We have introduced the mathematical problem at the core of this work with the objective of reviewing some of the stability properties of its *exact* solutions and to justify the *design constraints* imposed on the numerical schemes proposed in the thesis. The most important concepts introduced can be summarized as follows:

- Classical (pointwise) solutions to the differential problems can exist in general only in a bounded portion of the domain. The more appropriate and general concept of *weak solution* has been introduced;
- Existence and uniqueness of the solutions of the continuous problem need (or imply) the existence of some vanishing dissipative phenomenon. In the linear

case, this is measured by energy inequalities bounding the evolution of the energy (L^2 norm) of the solution or requiring this quantity to be non-increasing. In the nonlinear case, similar arguments lead to the existence of convex entropy pairs so that unique solutions are subject to an entropy stability statement;

- The unknowns of the problem often respect an L^∞ stability criterion, the maximum principle. This principle can be derived in a rigorous manner for scalar equations, while for systems the non-oscillatory character of the solution has been justified more heuristically. Even so, it appears reasonable to design methods for the approximation of weak solutions which produce (L^∞ -)stable solutions with no spurious oscillations in presence of discontinuities;

Chapter 3

Discrete approximation: grid, geometry and unknowns

The thesis presents and analyzes numerical discretization techniques for the solution of the model problem (2.1) on unstructured meshes. Before introducing the schemes considered in this work, this chapter illustrates how the spatial and temporal domains are discretized. In particular, throughout the manuscript the expressions irregular unstructured triangulation and structured triangulation will be often used. To clarify the difference between these two types of meshes used in the numerical experiments, we show here, *once and for all*, the mesh topologies we refer to.

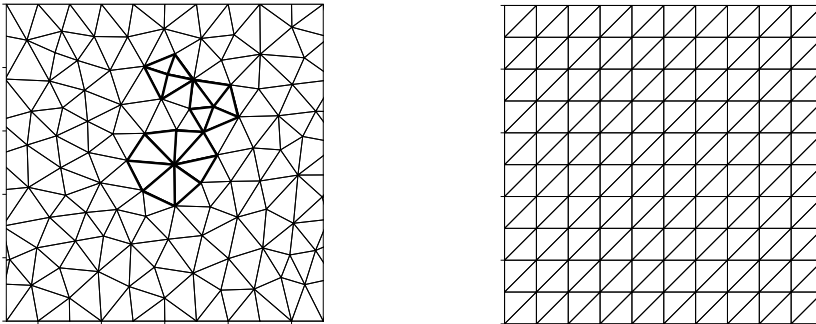


Figure 3.1: Unstructured (left) and structured (right) triangulation

Independently on the geometry of the problem at hand, whenever we refer to an unstructured discretization, we refer to a grid with the topology on the left on figure 3.1, obtained using a basic Weatherhill algorithm [17, 16]. These meshes have a roughly isotropic character, as far as size and angles are concerned. However, they present a somewhat random irregularity in the nodal connectivity, and consequently element

sizes and angles can vary considerably in a local neighborhood. This is well depicted in the left picture, where we have highlighted regions in which the number of neighbors of a node changes from four to eight. Structured triangulations are instead obtained simply by cutting into triangles meshes composed of squares. This is done always using the right-running diagonals, as on the right on figure 3.1. Except for one computation, discussed in chapter 10, we have never used any type of local refinement.

3.1 Mesh geometry

Consider a discretization of the spatial domain Ω composed by non overlapping triangular elements. We will denote the grid by \mathcal{T}_h , h being a reference element length, which is, for the grids used in this thesis, the (constant) mesh spacing on the boundary of Ω . We denote by E the generic triangle in \mathcal{T}_h and by $|E|$ its area. For all the grids considered here, the following *regularity* is assumed

$$0 < C_1 \leq \sup_{E \in \mathcal{T}_h} \frac{h^2}{|E|} \leq C_2 < \infty, \quad (3.1)$$

for finite positive constants C_1, C_2 . This corresponds to the fact that no *vanishing area* elements are present, as well as no very acute (or obtuse) angles. Given a node j in an element E , \vec{n}_j denotes the inward pointing vector normal to the edge of E opposite to j , scaled by the length of the edge (left on figure 3.2).

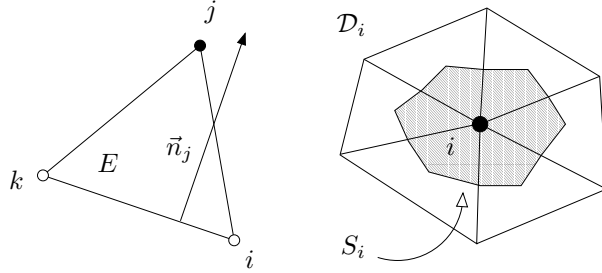


Figure 3.2: Median dual cell S_i and nodal normal \vec{n}_j

Note that since E has a closed boundary one has

$$\sum_{j \in E} \vec{n}_j = 0 \quad (3.2)$$

When no confusion is generated, we will locally number as $(1, 2, 3)$ the nodes of the generic triangle. For every node i in the mesh, \mathcal{D}_i denotes the subset of triangles containing i . By abuse of notation, we will say that $j \in \mathcal{D}_i$ if node j belongs to an element $E \in \mathcal{D}_i$. We then denote by S_i the median dual cell obtained by joining the gravity centers of the triangles in \mathcal{D}_i with the midpoints of the edges meeting in i , as

illustrated on the right on figure 3.2. The area of S_i is

$$|S_i| = \sum_{E \in \mathcal{D}_i} \frac{|E|}{3}, \quad (3.3)$$

We use the notation χ_S , $S \subset \Omega$, to denote the characteristic function of a subset S :

$$\chi_S(x, y) = \begin{cases} 1 & \text{if } (x, y) \in S \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

The temporal domain $[0, t_f]$ is subdivided by a sequence of M discrete time levels $\{t^1 = 0, t^2, \dots, t^n, t^{n+1}, \dots, t^M = t_f\}$. The schemes we consider allow, known the solution at a certain time t^n , to compute its approximation at time t^{n+1} . As a consequence, in most of the following chapters we focus our attention on the generic space-time slab $\Omega \times [t^n, t^{n+1}]$. The *time-width* of the slab is given by the time-step

$$\Delta t = t^{n+1} - t^n. \quad (3.5)$$

3.2 Variable and flux approximation

Once the spatial and temporal domain have been discretized, we introduce a discrete representation of the unknowns. The schemes developed in this thesis are based on a *continuous* approximation of the unknowns. In both the linear and the nonlinear case, this representation is constructed starting from the knowledge of the nodal values of the primary unknowns. It is perhaps best to distinguish between linear and nonlinear problems, since in the latter case different possibilities can (and will) be exploited. In particular, in this thesis we refer to the set of *primary* unknowns, as to the variables whose representation on the mesh is analytically known.

3.2.1 Discrete approximation of the unknowns: linear case

In the linear case, that is in the case of the scalar advection equation (2.7) or of the linear system (2.22), we will use as primary unknowns the variables u and \mathbf{u} respectively. In the following we refer always to the vector unknowns \mathbf{u} and, if not stated otherwise, it is assumed that the scalar case is formally obtained by replacing the \mathbf{u} vector by the scalar u . Let $\{\psi_i\}_{i \in \mathcal{T}_h}$ denote the continuous piecewise linear basis functions typically used in P^1 Finite Element methods, respecting

$$\psi_i(x_j, y_j) = \delta_{ij} \quad \forall i, j \in \mathcal{T}_h, \quad \nabla \psi_i|_E = \frac{\vec{n}_i}{2|E|}, \quad \sum_{j \in E} \psi_j = 1 \quad \forall E \in \mathcal{T}_h \quad (3.6)$$

with δ_{ij} Kronecker's delta. Given the nodal values $\{\mathbf{u}_i(t) = \mathbf{u}(x_i, y_i, t)\}_{i \in \mathcal{T}_h}$, we introduce the following continuous numerical approximation of \mathbf{u} in space

$$\mathbf{u}_h(x, y, t) = \sum_{i \in \mathcal{T}_h} \psi_i(x, y) \mathbf{u}_i(t). \quad (3.7)$$

The discrete representation of the initial solution reads

$$\mathbf{u}_h^0(x, y) = \sum_{i \in \mathcal{T}_h} \psi_i(x, y) \mathbf{u}_i^0 = \sum_{i \in \mathcal{T}_h} \psi_i(x, y) \mathbf{u}_0(x_i, y_i) . \quad (3.8)$$

For time-dependent computations, we will instead need a second-order accurate discrete representation on the space-time slab $\Omega \times [t^n, t^{n+1}]$. In this thesis, for linear problems, this representation is given by

$$\mathbf{u}^h(x, y, t) = \frac{t - t^n}{\Delta t} \mathbf{u}^{n+1} + \frac{t^{n+1} - t}{\Delta t} \mathbf{u}^n \quad \text{with} \quad \begin{cases} \mathbf{u}^n &= \mathbf{u}_h(x, y, t^n) \\ \mathbf{u}^{n+1} &= \mathbf{u}_h(x, y, t^{n+1}) \end{cases} \quad (3.9)$$

3.2.2 Discrete approximation of the unknowns: nonlinear case

In the nonlinear case, we need a discrete representation of both the unknowns and of the fluxes. In this case, and in particular for systems, different choices are possible. We denote by \mathbf{w} the generic set of primary unknowns. For example, in the case of nonlinear systems, we have at least the two possible choices $\mathbf{w} = \mathbf{u}$, the vector of conserved variables, and $\mathbf{w} = \mathbf{v}$, the entropy variables. Other choices are possible and will be presented in the text when and if needed. Moreover, by analogy with the system case, also for a scalar conservation law, we define the entropy variable

$$v = \frac{d\mathcal{H}(u)}{du} \quad (3.10)$$

for a given entropy pair $(\mathcal{H}(u), \mathcal{G}(u))$. As in the linear case, we introduce the representation of \mathbf{w} in space

$$\mathbf{w}_h(x, y, t) = \sum_{i \in \mathcal{T}_h} \psi_i(x, y) \mathbf{w}_i(t) . \quad (3.11)$$

It is intended, that the other variables are obtained as $\mathbf{u}_h = \mathbf{u}(\mathbf{w}_h)$. As in the linear case, the discrete initial solution is taken to be

$$\mathbf{w}_h^0(x, y) = \sum_{i \in \mathcal{T}_h} \psi_i \mathbf{w}_i^0 = \sum_{i \in \mathcal{T}_h} \psi_i \mathbf{w}(\mathbf{u}_0(x_i, y_i)) . \quad (3.12)$$

In time-dependent computations, we will use, on the space-time slab $\Omega \times [t^n, t^{n+1}]$, the second-order approximation

$$\mathbf{w}^h = \frac{t - t^n}{\Delta t} \mathbf{w}^{n+1} + \frac{t^{n+1} - t}{\Delta t} \mathbf{w}^n \quad \text{with} \quad \begin{cases} \mathbf{w}^n &= \mathbf{w}_h(x, y, t^n) \\ \mathbf{w}^{n+1} &= \mathbf{w}_h(x, y, t^{n+1}) \end{cases} \quad (3.13)$$

The other variables are obtained as $\mathbf{u}^h = \mathbf{u}(\mathbf{w}^h)$. The scalar case is obtained as the particular case $\mathbf{u} = u \in \mathbb{R}$. For nonlinear problems we also need an approximation of the fluxes. Generally, it will be assumed that $\mathcal{F}_h = \mathcal{F}(\mathbf{w}_h)$. However, it is useful to also introduce the following piecewise linear in time representation [8]

$$\mathcal{F}^h = \frac{t - t^n}{\Delta t} \mathcal{F}(\mathbf{w}^{n+1}) + \frac{t^{n+1} - t}{\Delta t} \mathcal{F}(\mathbf{w}^n) . \quad (3.14)$$

Both in the linear and in the nonlinear case, we denote by $\mathcal{S}_h(x, y)$ the discrete approximation of the source term, given by

$$\mathcal{S}_h = \sum_{i \in \mathcal{T}_h} \psi_i \mathcal{S}_i \quad (3.15)$$

3.3 Flux Jacobians

The schemes we will present in this thesis make use of information contained in the eigenstructure of the Jacobians of the fluxes. For completeness, we introduce the related labeling in this chapter. Using the notation of (2.4), on each element E we define the set of matrices

$$K_j = \frac{1}{2} K(\vec{n}_j, \bar{\mathbf{u}}), \quad (3.16)$$

where $\bar{\mathbf{u}}$ is an average value of \mathbf{u} over E that we leave, for the moment, unspecified. Note that, for simplicity of notation, we omit to use a superscript E (or subscript E) referring to the element in which K_j is defined, this being always clear from the context. Relation (3.2) implies that

$$\sum_{j \in E} K_j = 0 \quad (3.17)$$

Similarly to what we have done in chapter 2, we also introduce the *multidimensional upwind parameters* (see equation (2.5))

$$K_j^\pm = R_j \Lambda_j^\pm (R_j)^{-1} = \frac{1}{2} K^\pm(\vec{n}_j, \bar{\mathbf{u}}), \quad (3.18)$$

and the absolute value matrix

$$|K_j| = \frac{1}{2} |K(\vec{n}_j, \bar{\mathbf{u}})|. \quad (3.19)$$

Note that the analytical form of the Jacobians also depends on the choice of the primary variables. Unless stated otherwise, we will always keep the notation introduced here, also if the K_j matrix has been computed using flux Jacobians with respect to variables different from the conserved ones. Note also that (3.17) implies

$$\sum_{j \in E} K_j^+ = - \sum_{j \in E} K_j^- = \frac{1}{2} \sum_{j \in E} |K_j| \quad (3.20)$$

We introduce an additional set of Jacobians which will be useful in the presentation of the space-time schemes of [51]. In particular, denoting the $m \times m$ identity matrix by \mathbf{I} , we first define the following matrices

$$\tilde{K}_j = \frac{\Delta t}{2} K_j + \frac{|E|}{3} \mathbf{I}, \quad \hat{K}_j = \frac{\Delta t}{2} K_j - \frac{|E|}{3} \mathbf{I}. \quad (3.21)$$

The relation between the \tilde{K}_j and \hat{K}_j matrices and the Jacobians of the space-time flux $(\mathcal{F}, \mathbf{u})$ can be found in [51] and will be discussed later. We note that

$$\sum_{j \in E} (\tilde{K}_j + \hat{K}_j) = 0 \quad (3.22)$$

It can be easily verified that these matrices share with the K_j matrices (3.16) the same right and left eigenvectors and one can write

$$\tilde{K}_j = R_j \tilde{\Lambda}_j (R_j)^{-1}, \quad \hat{K}_j = R_j \hat{\Lambda}_j (R_j)^{-1} \quad (3.23)$$

with

$$\tilde{\Lambda}_j = \frac{\Delta t}{2} \Lambda_j + \frac{|E|}{3} \mathbf{I}, \quad \hat{\Lambda}_j = \frac{\Delta t}{2} \Lambda_j - \frac{|E|}{3} \mathbf{I} \quad (3.24)$$

We define then the *space-time multidimensional upwind parameters* [51]

$$\tilde{K}_j^\pm = R_j \tilde{\Lambda}_j^\pm (R_j)^{-1}, \quad \hat{K}_j^\pm = R_j \hat{\Lambda}_j^\pm (R_j)^{-1} \quad (3.25)$$

Due to (3.22), we can write the relation

$$\sum_{j \in E} (\tilde{K}_j^+ + \hat{K}_j^+) = - \sum_{j \in E} (\tilde{K}_j^- + \hat{K}_j^-) = \frac{1}{2} \sum_{j \in E} (|\tilde{K}_j| + |\hat{K}_j|) \quad (3.26)$$

3.3.1 Scalar Jacobians

In the scalar case, the flux Jacobians are vectors in \mathbb{R}^2 . We define the analog of (3.16) and of its positive and negative parts as

$$k_j^+ = \max(0, k_j), \quad k_j^- = \min(0, k_j); \quad k_j = \frac{\vec{a}(\bar{u}) \cdot \vec{n}_j}{2}, \quad (3.27)$$

with \vec{a} as in (2.16) or, in the case of the advection equation (2.7), given. The state \bar{u} is an average of u_h over the element whose properties we leave unspecified for the moment. Relations (3.17) and (3.20) are true also in the scalar case, with the obvious changes in notation.

For the space-time multidimensional upwind parameters one has in the scalar case

$$\tilde{k}_j^+ = \max(0, \tilde{k}_j), \quad \tilde{k}_j^- = \min(0, \tilde{k}_j) \quad \text{with} \quad \tilde{k}_j = \frac{\Delta t}{2} k_j + \frac{|E|}{3} \quad (3.28)$$

and

$$\hat{k}_j^+ = \max(0, \hat{k}_j), \quad \hat{k}_j^- = \min(0, \hat{k}_j) \quad \text{with} \quad \hat{k}_j = \frac{\Delta t}{2} k_j - \frac{|E|}{3} \quad (3.29)$$

Relations (3.22) and (3.26) apply also in this case, with obvious changes in notation.

Chapter 4

Prototype compact discrete approximation for steady advection

This chapter is devoted to the derivation of discrete analogs of the maximum principle and energy stability criteria introduced in chapter 2 for exact solutions of the linear advection equation. This will be done for a general abstract prototype encompassing the \mathcal{FS} schemes object of the thesis, and other methods as well. In particular, here we consider *cell-vertex* compact discretizations that, when computing steady-state solutions of the advection equation (2.7), and neglecting terms arising from the boundary conditions, can be recast in the following abstract form:

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \phi_i^E = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i - u_j), \quad \forall i \in \mathcal{T}_h \quad (4.1)$$

with $u_i(t=0) = u_0(x_i, y_i)$. The abstract scheme (4.1) is the most compact discretization one can have, since it only involves the *nearest neighboring* nodes of node i . We will see that this representation is particularly well suited for the \mathcal{RD} schemes developed in the thesis. However, we will also show that it encompasses linear \mathcal{FE} schemes and some first-order \mathcal{FV} schemes. A precise definition of a \mathcal{RD} scheme will be instead given in the next chapter. For the moment, we only assume that (4.1) is consistent with the advection equation (2.7):

Assumption (Consistency). *It is possible to find a consistent approximation of the unknown $u_h(x, y)$, or equivalently of the flux $\mathcal{F}_h(x, y) = (\bar{a}u)_h(x, y)$, such that scheme (4.1) verifies*

$$\sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} \phi_i^E = \oint_{\partial\Omega} \mathcal{F}_h \cdot \hat{n} \, dl, \quad (4.2)$$

with \hat{n} the exterior unit normal to $\partial\Omega$.

Moreover, without loss of generality, we will assume that the consistency requirement is verified on subsets of Ω as well, and in particular that

Assumption (Local Consistency). *It is possible to find a consistent approximation of the unknown $u_h(x, y)$, or equivalently of the flux $\mathcal{F}_h(x, y) = (\vec{a}u)_h(x, y)$, such that $\forall E \in \mathcal{T}_h$ scheme (4.1) verifies*

$$\sum_{j \in E} \phi_j^E = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n}_E \, dl, \quad (4.3)$$

with \hat{n}_E the exterior unit normal to ∂E .

We recall that (4.1) does not take into account terms arising from the boundary conditions, which are neglected in the analysis of this chapter. As we will show in the next chapter, local consistency is verified by all the \mathcal{FE} and \mathcal{FV} schemes represented by (4.1) and, by construction, by \mathcal{RD} schemes. We also mention that, if the local consistency (4.3) is verified, under a continuity hypothesis on the ϕ_i^E s, and assuming that the consistent approximation of the flux \mathcal{F}_h is continuous across triangle edges, it can be proved [5, 6, 9] that

Theorem (Lax-Wendroff theorem for prototype scheme). *Given bounded initial data $u_0 \in L^\infty(\mathbb{R}^2)$, a square integrable function $u \in L^2(\mathbb{R}^2 \times \mathbb{R}^+)$, and a constant C depending on u_0 and u such that the approximation $u_h(x, y, t)$ obtained from (4.1) verifies*

$$\sup_h \sup_{(x, y, t)} |u_h| \leq C \quad \lim_{h \rightarrow 0} \|u_h - u\|_{L^2_{loc}(\mathbb{R}^2 \times \mathbb{R}^+)} = 0,$$

then u is a weak solution of the problem, in the sense of (2.17).

As we will see in the next chapter, the hypotheses at the basis of this result apply to \mathcal{RD} schemes as well as to \mathcal{FE} and to \mathcal{FV} schemes that can be recast in form (4.1). In this chapter, we shall recall conditions under which the solutions obtained with this abstract scheme respect discrete analogs of the maximum principle (2.10) and of the energy stability condition (2.12).

4.1 Positive cell-vertex schemes on unstructured grids

In this section we consider the design of monotone schemes on unstructured meshes. Conditions are given on the c_{ij}^E coefficients in (4.1) in order to be able to prove that the numerical solution obtained respects discrete analogs of the maximum principle (2.10). The first part of the analysis is a re-adaptation to the case of our cell-vertex prototype (4.1) of the theory of *positive coefficient schemes* on unstructured meshes, discussed for \mathcal{FV} schemes in [20, 25] (see also [161]). We then extend the theory to implicit two-level schemes and to non-homogeneous problems where the source term does not depend on the solution.

The analysis starts with a very important property of (4.1): the so-called *Local Extremum Diminishing* (LED) property:

Proposition 4.1.1 (LED property). *The prototype scheme (4.1) is Local Extremum Diminishing (LED), i.e. in the numerical solution local maxima are non-increasing and local minima are non-decreasing, if*

$$\tilde{c}_{ij} = \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} c_{ij}^E \geq 0, \quad \forall j \in \mathcal{D}_i, j \neq i \text{ and } \forall i \in \mathcal{T}_h \quad (4.4)$$

Proof. From property (4.4) it follows that

$$\begin{aligned} \frac{du_i}{dt} &= -\frac{1}{|S_i|} \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i - u_j) = \\ &= -\frac{1}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \left(\sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} c_{ij}^E \right) (u_i - u_j) = -\frac{1}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} (u_i - u_j) \end{aligned}$$

is negative or zero if u_i is a local maximum ($u_i \geq u_j$) and it is positive or zero if u_i is a local minimum ($u_i \leq u_j$), hence the result. \square

The LED property guarantees that local extrema are kept bounded by the numerical scheme. A stronger requirement, often used in the construction of nonlinear schemes, is trivially obtained by asking each c_{ij}^E to be positive on every element E , leading to a *sub-element LED* property:

Corollary 4.1.2 (Sub-element LED). *Scheme (4.1) is LED if*

$$c_{ij}^E \geq 0 \quad \forall j \in E \text{ and } \forall E \in \mathcal{D}_i. \quad (4.5)$$

In order to obtain an estimate on the discrete solution, the LED condition is not enough and fully discrete versions of (4.1) need to be considered. In this work, three types of two-level explicit and implicit time discretizations are analyzed: explicit (forward) Euler (FE), implicit (backward) Euler (BE), Crank-Nicholson (\mathcal{CN}) and trapezium rule. For linear problems, the last two are equivalent. For clarity, we analyze the explicit scheme first. This part of the theory is a re-adaptation to our cell-vertex framework of the analysis reported in [20, 25]. We then extend the theory to the implicit case and to the non-homogeneous one.

4.1.1 Discrete maximum principle: explicit case

Consider the fully discrete version of (4.1) obtained with the explicit FE scheme:

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|S_i|} \sum_{E \in \mathcal{D}_i} \phi_i^{\text{FE}} = u_i^n - \frac{\Delta t}{|S_i|} \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i^n - u_j^n). \quad (4.6)$$

We have the following result.

Proposition 4.1.3 (Positivity-Local Discrete Maximum Principle). *The space-time discrete analog of (2.7) in the time interval $[t^n, t^{n+1}]$ represented by (4.6), verifies the local space-time maximum principle*

$$\tilde{u}_i^n \leq u_i^{n+1} \leq \tilde{U}_i^n, \quad (4.7)$$

with

$$\tilde{u}_i^n = \min_{j \in \mathcal{D}_i} u_j^n, \quad \tilde{U}_i^n = \max_{j \in \mathcal{D}_i} u_j^n, \quad (4.8)$$

if the LED condition (4.4) holds and under the time-step restriction

$$\Delta t \leq \frac{|S_i|}{\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}} \quad \forall i \in \mathcal{T}_h \quad (4.9)$$

Proof. Rewriting (4.6) as

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\Delta t}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} (u_i^n - u_j^n) = \\ &\left(1 - \frac{\Delta t}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} \right) u_i^n + \frac{\Delta t}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^n = \bar{c}_{ii} u_i^n + \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \bar{c}_{ij} u_j^n = \sum_{j \in \mathcal{D}_i} \bar{c}_{ij} u_j^n, \end{aligned}$$

properties (4.4) and (4.9) guarantee that for all i and j , $\bar{c}_{ij} \geq 0$, hence:

$$\left(\sum_{j \in \mathcal{D}_i} \bar{c}_{ij} \right) \tilde{u}_i^n \leq u_i^{n+1} \leq \left(\sum_{j \in \mathcal{D}_i} \bar{c}_{ij} \right) \tilde{U}_i^n.$$

Using the fact that, trivially,

$$\sum_{j \in \mathcal{D}_i} \bar{c}_{ij} = 1, \quad (4.10)$$

we obtain the proof. \square

A scheme of the form (4.6) respecting proposition 4.1.3 is said to be *positive*. As done for the LED property, we introduce a local form of positivity, which will be useful later for the construction of compact nonlinear schemes. In particular, using (3.3) we can rewrite (4.1) as:

$$|S_i| u_i^{n+1} = \sum_{E \in \mathcal{D}_i} \left(\frac{|E|}{3} u_i^n - \Delta t \phi_i^{\text{FE}} \right) \Rightarrow u_i^{n+1} = \sum_{E \in \mathcal{D}_i} \frac{|E|}{3|S_i|} u_i^E, \quad (4.11)$$

with

$$u_i^E = u_i^n - \frac{3\Delta t}{|E|} \phi_i^{\text{FE}} = u_i^n - \frac{3\Delta t}{|E|} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i^n - u_j^n). \quad (4.12)$$

With this notation we introduce the following result.

Proposition 4.1.4 (Local Positivity-Discrete maximum principle). *The space-time discrete analog of (2.7) given by (4.6) exhibits the local maximum discrete principle (4.7), if it verifies the sub-element LED condition (4.5) and under the time-step restriction*

$$\Delta t \leq \frac{|E|}{3 \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E} \quad \forall i \in \mathcal{T}_h \quad (4.13)$$

Proof. One easily checks that the hypotheses of proposition 4.1.3 are verified under the assumptions made here. \square

A scheme respecting proposition 4.1.4 is said to be *locally positive*. Last proposition states that local positivity implies positivity. Two important consequences have to be recalled. The first is that, thanks to the positivity of the \tilde{c}_{ij} coefficients implied by condition (4.4), we have [25, 20]:

Proposition 4.1.5 (Steady-state discrete maximum principle). *Under the hypothesis that the \tilde{c}_{ij} coefficients in (4.4) are all positive, the steady limit of (4.6) verifies the local maximum principle in space given by:*

$$\min_{\substack{j \in \mathcal{D}_i \\ j \neq i}} u_j^* \leq u_i^* \leq \max_{\substack{j \in \mathcal{D}_i \\ j \neq i}} u_j^* \quad (4.14)$$

where the superscript $*$ denotes the steady limit

$$u_j^* = \lim_{n \rightarrow \infty} u_j^n .$$

The second and more important consequence is that the solution respects at all times a discrete analog of (2.10) [25, 20]:

Theorem 4.1.6 (L^∞ -stability). *If the hypotheses of proposition 4.1.3 are verified in all the time slabs $\{[t^n, t^{n+1}]\}_{n=0, \dots, M-1}$, then scheme (4.6) is L^∞ -stable and the following bounds hold for its numerical solution:*

$$\min_{i \in \mathcal{T}_h} u_i^0 \leq u_j^n \leq \max_{i \in \mathcal{T}_h} u_i^0, \quad \forall i \in \mathcal{T}_h, n \in [1, M] . \quad (4.15)$$

4.1.2 Discrete maximum principle: implicit case

We consider now the case in which (4.1) is integrated in time using an implicit scheme. This part of the analysis will be very important in the construction of monotone schemes for time-dependent computations. In particular, in this section we consider the analysis of the *implicit backward Euler scheme* (BE)

$$|S_i|(u_i^{n+1} - u_i^n) = -\Delta t \sum_{E \in \mathcal{D}_i} \phi_i^{\text{BE}} = -\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i^{n+1} - u_j^{n+1}), \quad (4.16)$$

and of the *implicit Crank-Nicholson* (\mathcal{CN}) scheme

$$|S_i|(u_i^{n+1} - u_i^n) = -\Delta t \sum_{E \in \mathcal{D}_i} \phi_i^{\mathcal{CN}} = -\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \left(\frac{u_i^{n+1} + u_i^n}{2} - \frac{u_j^{n+1} + u_j^n}{2} \right) \quad (4.17)$$

Note that in the linear case considered here, the \mathcal{CN} scheme is equivalent to the so-called *trapezium time integration scheme*

$$|S_i|(u_i^{n+1} - u_i^n) = -\Delta t \sum_{E \in \mathcal{D}_i} \frac{\phi_i^{\text{FE}} + \phi_i^{\text{BE}}}{2}. \quad (4.18)$$

Denoting by U^n and U^{n+1} the arrays containing the nodal values of u at time t^n and t^{n+1} respectively, the implicit schemes presented can all be recast in the form:

$$\mathcal{A}U^{n+1} = \mathcal{B}U^n, \quad (4.19)$$

where the \mathcal{A} and \mathcal{B} matrices are sparse with a fill-in pattern given by the connectivity graph of the grid¹. The entries of these matrices depend on the c_{ij}^E coefficients, on the time-step and on S_i . These entries can be expressed in a general unified form introducing the θ -scheme [99]:

$$|S_i|(u_i^{n+1} - u_i^n) = -\Delta t \sum_{E \in \mathcal{D}_i} ((1 - \theta)\phi_i^{\text{FE}} + \theta\phi_i^{\text{BE}}). \quad (4.20)$$

The θ -scheme, can be recast as in (4.19) with \mathcal{A} and \mathcal{B} given by

$$\begin{aligned} \mathcal{A}_{ii} &= |S_i| + \theta\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, & \mathcal{A}_{ij} &= -\theta\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \\ \mathcal{B}_{ii} &= |S_i| - (1 - \theta)\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, & \mathcal{B}_{ij} &= (1 - \theta)\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \end{aligned} \quad (4.21)$$

Clearly, the BE scheme is obtained for $\theta = 1$ and the \mathcal{CN} scheme corresponds to the choice $\theta = 1/2$. Now we can prove

Proposition 4.1.7 (Positivity- Discrete Maximum Principle). *The space-time discrete analog of (2.7) in the time interval $[t^n, t^{n+1}]$ represented by the θ -scheme (4.20), verifies the global discrete space-time maximum principle*

$$u_{\min}^n = \min_{j \in \mathcal{T}_h} u_j^n \leq u_i^{n+1} \leq \max_{j \in \mathcal{T}_h} u_j^n = u_{\max}^n, \quad (4.22)$$

and the local discrete space-time maximum principle given by

$$\bar{u}_i = \min \left\{ u_i^n, \min_{\substack{j \in \mathcal{D}_i \\ j \neq i}} (u_j^n, u_j^{n+1}) \right\} \leq u_i^{n+1} \leq \max \left\{ u_i^n, \max_{\substack{j \in \mathcal{D}_i \\ j \neq i}} (u_j^n, u_j^{n+1}) \right\} = \bar{U}_i \quad (4.23)$$

¹i.e in general \mathcal{A}_{lm} (and \mathcal{B}_{lm}) is non-zero only if there is at least one element $E \in \mathcal{T}_h$ such that both nodes l and m belong to E

if the LED condition (4.4) holds and under the time-step restriction

$$|S_i| - (1 - \theta)\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \geq 0 \quad \forall i \in \mathcal{T}_h \quad (4.24)$$

In particular, the BE scheme (4.16) verifies (4.22) and (4.23) $\forall \Delta t > 0$, while the time-step restriction of the CN scheme is twice less severe than the one guaranteeing the positivity of the FE scheme, equation (4.9).

Proof. The proof is obtained by noting that the LED condition (4.4) guarantees that $\mathcal{A}_{ii} \geq 0$ and $\mathcal{A}_{ij} \leq 0 \forall j \neq i$ independently on Δt . Moreover, \mathcal{A} is diagonally dominant since

$$|\mathcal{A}_{ii}| - \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} |\mathcal{A}_{ij}| = |S_i| > 0.$$

Hence, \mathcal{A} is an \mathcal{M} -matrix and is diagonally dominant. This implies that \mathcal{A} is invertible and \mathcal{A}^{-1} is positive [108]: $\mathcal{A}_{ij}^{-1} \geq 0 \forall i, j$. Consider now the array U_{\min} having the same length of U^n and U^{n+1} but with elements all equal to u_{\min}^n . Thanks to the time-step restriction (4.24), we have $\mathcal{B}_{ij} \geq 0 \forall i, j$, hence

$$(\mathcal{B}U^n)_i \geq (\mathcal{B}U_{\min})_i \quad \forall i \in \mathcal{T}_h$$

since $u_i^n \geq u_{\min}^n \forall i \in \mathcal{T}_h$. Moreover

$$(\mathcal{B}U_{\min})_i = \sum_{j \in \mathcal{D}_i} \mathcal{B}_{ij} u_{\min}^n = |S_i| u_{\min}^n = \sum_{j \in \mathcal{D}_i} \mathcal{A}_{ij} u_{\min}^n = (\mathcal{A}U_{\min})_i$$

Since $\mathcal{A}U^{n+1} = \mathcal{B}U^n$, this shows that

$$(\mathcal{A}U^{n+1})_i \geq (\mathcal{A}U_{\min})_i \quad \forall i \in \mathcal{T}_h$$

The positivity of $\mathcal{A}^{-1} \geq 0$ implies the left inequality in (4.22). The right inequality is obtained considering the array U_{\max} with entries all equal to u_{\max}^n and proceeding in a similar way.

As in the case of the steady-state maximum principle (4.14), the local bounds are a consequence of the positivity of the \tilde{c}_{ij} coefficients. In fact, given U^{n+1} , on has for a node i :

$$\begin{aligned} \mathcal{A}_{ii} u_i^{n+1} + \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \mathcal{A}_{ij} u_j^{n+1} &= \left(|S_i| + \theta \Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} \right) u_i^{n+1} - \theta \Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^{n+1} = \\ \mathcal{B}_{ii} u_i^n + \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \mathcal{B}_{ij} u_j^n &= \left(|S_i| - (1 - \theta) \Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} \right) u_i^{n+1} + (1 - \theta) \Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^n \end{aligned}$$

Thanks to the positivity of the \tilde{c}_{ij} s and to the time-step restriction (4.24), and using the definition of \bar{U}_i (equation (4.23)), one has

$$\begin{aligned} \left(|S_i| + \theta\Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}\right) u_i^{n+1} = \\ \left(|S_i| - (1-\theta)\Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}\right) u_i^{n+1} + (1-\theta)\Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^n + \theta\Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^{n+1} \leq \\ \left(|S_i| + \theta\Delta t \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}\right) \bar{U}_i \end{aligned} \quad (4.25)$$

which gives the right bound in (4.23). The left bound is obtained in a similar way. In the case of the BE scheme one has $\theta = 1$, hence (4.24) is satisfied $\forall \Delta t > 0$. Setting $\theta = 1/2$ in (4.24), we obtain for the CN scheme

$$\Delta t \leq 2 \frac{|S_i|}{\sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E} = 2 \frac{|S_i|}{\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}},$$

where the left hand side is exactly twice as large as the one in (4.9). \square

An implicit scheme respecting proposition 4.1.7 will be said to be *positive*. Next, we note that the components of the \mathcal{A} and \mathcal{B} matrices can be decomposed as a sum of local contributions:

$$\mathcal{A} = \sum_{E \in \mathcal{T}_h} \mathcal{A}^E, \quad \mathcal{B} = \sum_{E \in \mathcal{T}_h} \mathcal{B}^E$$

where $\mathcal{A}_{ij}, \mathcal{B}_{ij} = 0 \forall i, j \notin E$, and for $i, j \in E$ one has

$$\begin{aligned} \mathcal{A}_{ii}^E &= \frac{|E|}{3} + \theta\Delta t \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, & \mathcal{A}_{ij}^E &= -\theta\Delta t \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \\ \mathcal{B}_{ii}^E &= \frac{|E|}{3} - (1-\theta)\Delta t \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, & \mathcal{B}_{ij}^E &= (1-\theta)\Delta t \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \end{aligned} \quad (4.26)$$

Trivially

Proposition 4.1.8 (Local Positivity - Discrete Maximum Principle). *The space-time discrete analog of (2.7) in the time interval $[t^n, t^{n+1}]$ represented by the θ -scheme (4.20), verifies the global space-time discrete maximum principle (4.22) and the local space-time discrete maximum principle (4.23) if the sub-element LED condition (4.5) holds and under the time-step restriction*

$$\frac{|E|}{3} - (1-\theta)\Delta t \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E \geq 0 \quad \forall i \in E \text{ and } \forall E \in \mathcal{T}_h \quad (4.27)$$

In particular, the BE scheme (4.16) verifies (4.22) and (4.23) $\forall \Delta t > 0$, while the time-step restriction of the CN scheme is twice less severe than the one guaranteeing the local positivity of the FE scheme, equation (4.13).

Proof. The sub-element LED condition implies the LED condition and (4.27) implies (4.24). Application of proposition 4.1.7 yields the result. As before, if the local LED condition holds, then the implicit BE scheme is positive $\forall \Delta t > 0$, while for the CN scheme we have the time-step restriction

$$\Delta t \leq 2 \frac{|E|}{3 \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E} \quad \forall i \in \mathcal{T}_h$$

which is precisely twice less strict than (4.13). \square

Implicit schemes verifying proposition 4.1.8 are said to be *locally positive*. Clearly, last proposition shows that local positivity implies positivity. It must be remarked that it seems quite disappointing that an implicit scheme must respect a time-step restriction of the same order as the one of the explicit FE scheme in order to preserve the L^∞ stability of the discretization. Unfortunately, it can be shown that, as far as high-order time-integration schemes are concerned, this strict limitation has a quite general character [27]. We will discuss this issue in some more detail later when presenting the space-time schemes. Here we add that, as in the case of the explicit scheme, the LED property implies that the θ -scheme satisfies the steady state maximum discrete principle of proposition 4.1.5 and moreover

Theorem 4.1.9 (L^∞ -stability). *If the hypotheses of proposition 4.1.7 are verified in all the time slabs $\{[t^n, t^{n+1}]\}_{n=0, \dots, M-1}$, then scheme (4.20) is L^∞ -stable and the numerical solution respects bounds (4.15).*

4.2 Energy stability

As seen in §2.1, the advection equation is characterized by a bound on the L^2 norm of its exact solutions: the energy. At the discrete level, this translates into a stability criterion: stable schemes are the ones for which the energy attains its maximum value at $t = 0$. Hence, the energy is *dissipated* by stable discretizations. In this section we give estimates for the evolution in time of the energy of the solution obtained by scheme (4.1). The analysis is partly inspired by [18]. We start by rewriting the prototype scheme in the compact vector form

$$D_{|S_i|} \frac{dU}{dt} = -CU, \quad (4.28)$$

where $D_{|S_i|}$ is the diagonal matrix of the median dual areas, U is the array containing the nodal values $u_i \forall i \in \mathcal{T}_h$ and the form of matrix C are given by the \tilde{c}_{ij} coefficients

of the LED condition (4.4):

$$c_{ii} = \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}, \quad c_{ij} = -\tilde{c}_{ij} = - \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} c_{ij}^E. \quad (4.29)$$

Introducing the discrete analog of the energy of the solution

$$\mathcal{E}_h = \frac{U^T D_{|S_i|} U}{2}, \quad (4.30)$$

the energy dissipation of the schemes can be characterized by analyzing

$$\frac{d\mathcal{E}_h}{dt} = -U^T \frac{\mathcal{C} + \mathcal{C}^T}{2} U = -U^T M^{\mathcal{E}_h} U. \quad (4.31)$$

We start giving the following definition.

Definition 4.2.1 (Energy stable scheme - semi-discrete form). *The prototype scheme in semi-discrete form (4.1) is energy stable if the symmetric matrix $M^{\mathcal{E}_h}$ is positive semi-definite, hence*

$$\frac{d\mathcal{E}_h}{dt} = -U^T M^{\mathcal{E}_h} U \leq 0.$$

It is common experience that schemes yielding monotone numerical solutions, such as LED and positive schemes, also exhibit a *dissipative behavior*, i.e. sharp profiles of the solution are smeared as if a viscous diffusion mechanism was present. In order to characterize our prototype scheme from the energy point of view, we look at the form of the $M^{\mathcal{E}_h}$ matrix. In particular, from (4.31) and (4.29) we have

$$M_{ii}^{\mathcal{E}_h} = \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}, \quad M_{ij}^{\mathcal{E}_h} = -\frac{1}{2} (\tilde{c}_{ij} + \tilde{c}_{ji}) = - \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} \frac{c_{ij}^E + c_{ji}^E}{2}.$$

For LED schemes $M^{\mathcal{E}_h}$ has positive entries on the diagonal and negative off-diagonal terms. However, this is not enough to ensure the positive semi-definiteness, unless the matrix is also irreducibly diagonally dominant [26]. In particular, some of the schemes we consider in this thesis can be characterized by the following property.

Proposition 4.2.2 (Energy stability of LED schemes - semi-discrete case). *A scheme of the form (4.1) verifying the LED condition (4.4) is energy stable in the sense of definition 4.2.1 if*

$$\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} = \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ji} \quad \forall i \in \tau_h \quad (4.32)$$

Proof. We start by recasting the quadratic form on the right hand side in (4.31) as

$$\begin{aligned} U^T M^{\mathcal{E}_h} U &= \sum_{i \in \mathcal{T}_h} \left\{ u_i \left(\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} \right) u_i - u_i \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \frac{\tilde{c}_{ij} + \tilde{c}_{ji}}{2} u_j \right\} = \\ &\quad \sum_{i \in \mathcal{T}_h} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} u_i \frac{\tilde{c}_{ij} + \tilde{c}_{ji}}{2} (u_i - u_j) + \sum_{i \in \mathcal{T}_h} u_i \left(\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \frac{\tilde{c}_{ij} - \tilde{c}_{ji}}{2} \right) u_i \end{aligned}$$

Rewriting the first term in the last expression as a sum involving mesh edges we get:

$$U^T M^{\mathcal{E}_h} U = \frac{1}{2} \sum_{\substack{i, j \in \mathcal{T}_h \\ \mathcal{D}_i \cap \mathcal{D}_j \neq \emptyset}} \overbrace{(u_i - u_j) \frac{\tilde{c}_{ij} + \tilde{c}_{ji}}{2} (u_i - u_j)}^{\geq 0} + \sum_{i \in \mathcal{T}_h} u_i \left(\sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \frac{\tilde{c}_{ij} - \tilde{c}_{ji}}{2} \right) u_i$$

Due to the satisfaction of the LED condition (4.4), the first sum is always non-negative. If $M^{\mathcal{E}_h}$ is diagonally dominant, then

$$|M_{ii}^{\mathcal{E}_h}| - \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} |M_{ij}^{\mathcal{E}_h}| = \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} - \frac{1}{2} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} (\tilde{c}_{ij} + \tilde{c}_{ji}) = \frac{1}{2} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} (\tilde{c}_{ij} - \tilde{c}_{ji}) \geq 0,$$

hence $U^T M^{\mathcal{E}_h} U \geq 0$ and the scheme is guaranteed to be stable. In particular, LED schemes respecting (4.32) are energy stable. \square

We recall that the cell-vertex prototype analyzed in this chapter does not take into account terms arising from the boundary conditions. Complete energy stability estimates will have to be derived by combining proposition 4.2.2 with these terms. Some examples showing how boundary conditions can be included in the analysis will be given in the next chapter. Here, instead, we will give a more local characterization of the stability of the schemes. We start recalling the following result [18]

Lemma 4.2.3 (Energy equivalence - Barth, 1996). *Two matrix operators \mathcal{L}_1 and \mathcal{L}_2 for which*

$$\mathcal{L}_1 U = \mathcal{L}_2 U$$

are energy equivalent.

Proof. Trivially: $U^T(\mathcal{L}_1 + \mathcal{L}_1^T)U = U^T(\mathcal{L}_1 U) + (\mathcal{L}_1 U)^T U = U^T(\mathcal{L}_2 U) + (\mathcal{L}_2 U)^T U$, and hence $U^T(\mathcal{L}_1 + \mathcal{L}_1^T)U = U^T(\mathcal{L}_2 + \mathcal{L}_2^T)U$ \square

Barth's lemma has two important consequences. The first is the following property.

Proposition 4.2.4. *Schemes of the form (4.1) respecting (4.32) are energy equivalent to the semi-discrete evolution schemes*

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} \left(c_{ij}^E (u_i - u_j) - \frac{1}{2} (c_{ij}^E - c_{ji}^E) u_i \right) \quad \forall i \in \mathcal{T}_h$$

with associated matrix energy operator $\overline{M}^{\mathcal{E}_h}$ with entries

$$\overline{M}_{ii}^{\mathcal{E}_h} = \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \frac{\tilde{c}_{ij} + \tilde{c}_{ji}}{2}, \quad \overline{M}_{ij}^{\mathcal{E}_h} = -\frac{\tilde{c}_{ij} + \tilde{c}_{ji}}{2}. \quad (4.33)$$

LED schemes are energy stable with respect to the equivalent energy operator $\overline{M}^{\mathcal{E}_h}$.

The energy equivalence lemma and proposition 4.2.4 finally allow to give a local characterization of the energy stability of the schemes. To do this, we use the fact that (4.1) is obtained as a sum of elemental contributions:

$$U^T D_{|S_i|} \frac{dU}{dt} = \sum_{i \in \mathcal{T}_h} u_i |S_i| \frac{du_i}{dt} = \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} u_i \frac{|E|}{3} \frac{du_i}{dt} = \sum_{E \in \mathcal{T}_h} \sum_{i \in E} u_i \frac{|E|}{3} \frac{du_i}{dt}$$

and finally

$$U^T D_{|S_i|} \frac{dU}{dt} = \sum_{E \in \mathcal{T}_h} U_E^T D_E \frac{dU_E}{dt} \quad (4.34)$$

with U_E the array containing the three nodal values of the unknown in the vertices of E and D_E the diagonal matrix of $|E|/3$. Using the local numbering (u_1, u_2, u_3) for these quantities, one easily checks that

$$\frac{|E|}{3} \frac{du_l}{dt} = -\phi_l^E, \quad l = 1, 2, 3 \implies U_E^T D_E \frac{dU_E}{dt} = -U_E^T \Phi_E$$

with $\Phi_E = [\phi_1^E, \phi_2^E, \phi_3^E]^T$. Dropping for clarity the sub and super-scripts E , we define the *local energy* on an element as

$$\mathcal{E} = \frac{U^T D U}{2}, \quad (4.35)$$

where we recall that in the last expression (and in the following ones) $U = U_E$ and $D = D_E$ (and similarly for all the other quantities). Next, we note that

$$\Phi = \mathcal{C}U$$

with

$$C_{ii} = \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, \quad \text{and} \quad C_{ij} = -c_{ij}^E \quad \forall i, j = 1, 2, 3$$

The evolution of the local energy is then governed by

$$\frac{d\mathcal{E}}{dt} = -U^T \frac{\mathcal{C} + \mathcal{C}^T}{2} U = -U^T M U, \quad (4.36)$$

where, trivially

$$M^{\mathcal{E}_h} = \sum_{E \in \mathcal{T}_h} M. \quad (4.37)$$

We define a locally energy stable scheme as follows.

Definition 4.2.5 (Locally Energy stable scheme - semi-discrete form). *The prototype scheme in semi-discrete form (4.1) is locally energy stable if the symmetric matrix M is positive semi-definite, hence*

$$\frac{d\mathcal{E}}{dt} = -U^T M U \leq 0 .$$

Due to (4.34) and (4.37), we have the following trivial proposition:

Proposition 4.2.6. *Local energy stability implies energy stability.*

Unfortunately, we are not able to characterize the local stability of a scheme. However, we can construct a local criterion to check the global stability of the discretization. First of all we note that the entries of the matrix M are

$$M_{ii} = \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E, \quad M_{ij} = -\frac{1}{2} (c_{ij}^E + c_{ji}^E),$$

hence for schemes respecting the sub-element LED condition (4.5), M has all positive diagonal elements and all non-positive off-diagonal terms. This can be used to prove the following property.

Proposition 4.2.7 (Energy stability and sub-element LED - semi-discrete case). *A scheme of the form (4.1) verifying the sub-element LED condition (4.5) and condition (4.32) is energy stable. Moreover, it is locally energy stable with respect to the equivalent energy operator $\overline{M}^{\mathcal{E}_h}$.*

Proof. The first assertion is a consequence of the fact that schemes respecting the sub-element LED condition are LED, hence they satisfy proposition 4.2.2. To obtain the second assertion, we note that:

$$\overline{M}^{\mathcal{E}_h} = \sum_{E \in \mathcal{T}_h} \overline{M}$$

with

$$\overline{M}_{ii} = \sum_{\substack{j \in E \\ j \neq i}} \frac{1}{2} (c_{ij}^E + c_{ji}^E), \quad \overline{M}_{ij} = -\frac{1}{2} (c_{ij}^E + c_{ji}^E) .$$

Due to the sub-element LED condition \overline{M} defines a non-negative quadratic form. In particular, the local energy balance reads

$$\frac{d\mathcal{E}}{dt} = -U^T \overline{M} U = -\frac{1}{2} \sum_{i,j \in E} (u_i - u_j) \frac{c_{ij}^E + c_{ji}^E}{2} (u_i - u_j) \leq 0 \quad \forall E \in \mathcal{T}_h ,$$

showing that the local LED condition and the diagonal dominance condition (4.32) imply the local stability of the scheme with respect to $\overline{M}^{\mathcal{E}_h}$. \square

4.2.1 Fully discrete case

The analysis of the previous section has allowed to characterize the *dissipative* character of the spatial discretizations represented by the prototype (4.1). However, the stability of the discretization has also to take into account the discretization of the time derivative. For the schemes considered here, this can be done in a general way noting that the θ -scheme (4.20) encompasses the explicit FE scheme ($\theta = 0$), the implicit BE scheme ($\theta = 1$) and the \mathcal{CN} scheme ($\theta = 1/2$). We have the following result.

Proposition 4.2.8 (Discrete energy stability - θ -scheme). *The family of schemes represented by the θ -scheme (4.20) verify the following fully discrete energy balance*

$$\mathcal{E}_h^{n+1} = \mathcal{E}_h^n - \Delta t \left(\theta U^{n+1} + (1 - \theta) U^n \right)^T M^{\mathcal{E}_h} \left(\theta U^{n+1} + (1 - \theta) U^n \right) - (2\theta - 1) \epsilon_h \quad (4.38)$$

with the discrete time energy production ϵ_h given by

$$\epsilon_h = \frac{1}{2} (U^{n+1} - U^n)^T D_{|S_i|} (U^{n+1} - U^n) \geq 0.$$

The time discretization has a stabilizing effect for $\theta > 1/2$ and a destabilizing effect for $\theta < 1/2$. In particular, the explicit FE time discretization has the maximum energy destabilizing character and the implicit BE scheme is the most stable. The \mathcal{CN} scheme is the only one preserving the dissipation properties of the spatial discretization. For this reason the \mathcal{CN} scheme is said to be energy conservative.

Proof. The proof reduces to showing that the balance (4.38) is true. The remaining assertions are trivially verified by analyzing the sign of the additional term in the balance, governed by the quantity $2\theta - 1$. The energy balance is easily obtained by first noting that

$$\theta u_i^{n+1} + (1 - \theta) u_i^n = \frac{u_i^{n+1} + u_i^n}{2} + (2\theta - 1) \frac{u_i^{n+1} - u_i^n}{2} \quad \forall i \in \mathcal{T}_h$$

Upon multiplication of (4.20) by $\theta u_i^{n+1} + (1 - \theta) u_i^n$, and summing the expression thus obtained to its transpose, we obtain the desired result. \square

Note that, as a consequence of the last proposition, while implicit schemes with $\theta > 1/2$ might stabilize space discretizations which, by themselves, are not energy stable, the use of the FE scheme (or in general of schemes with $\theta < 1/2$) might spoil the energy stability of the spatial discrete operator. These competitive effects can be controlled by changing the magnitude of the time-step. For energy stable space discretizations, one might then seek a limiting value of Δt for the time discretization guaranteeing the stability. This study, not undertaken here, can lead sometimes to time-steps constraints for energy stability which are stricter than the ones proved to yield the L^∞ stability of the space discretization (see *e.g.* [171]).

4.3 Stability: the inhomogeneous case

The analysis of the non-homogeneous case

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \mathcal{S}(x, y) \quad (4.39)$$

is considered here. In particular, we derive some L^∞ bounds on the discrete solution, while we give no discrete energy estimates for which one can refer for example to [152, 30]. As in chapter 2, we assume that $\mathcal{S}(x, y)$ is uniformly bounded

$$\sup_{\Omega} |\mathcal{S}(x, y)| < \infty .$$

We will consider schemes that, when used to discretize (4.39), can be recast in the form

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \phi_i^E = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i - u_j) + \sum_{E \in \mathcal{D}_i} \sum_{j \in E} |E| c_{ij}^{\mathcal{S}} \mathcal{S}_j , \quad (4.40)$$

with $\mathcal{S}_j = \mathcal{S}(x_j, y_j)$. The discretization of the source-term might be somehow dependent on the discrete advective operator. Hence the $c_{ij}^{\mathcal{S}}$ might depend on \vec{a} . As a consequence, we *will not* require that

$$\sum_{j \in E} c_{ij}^{\mathcal{S}} = 1 .$$

However, we do require some form of consistency of the discretization, without loss of generality expressed by

Assumption (Local Consistency - inhomogeneous case). *For a given scheme of the form (4.40), it is possible to find a consistent approximation of the unknown $u_h(x, y)$, or equivalently of the flux $\mathcal{F}_h(x, y) = (\vec{a}u)_h(x, y)$, and of the source term \mathcal{S}_h , such that $\forall E \in \mathcal{T}_h$*

$$\sum_{j \in E} \phi_j^E = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n}_E \, dl - \int_E \mathcal{S}_h \, dx \, dy , \quad (4.41)$$

with \hat{n}_E the unit normal to ∂E .

As a consequence we have that

$$\sum_{i \in E} \beta_i^{\mathcal{S}} = 1 \quad \text{with} \quad \beta_i^{\mathcal{S}} = \sum_{j \in E} c_{ij}^{\mathcal{S}} \quad \forall i \in E . \quad (4.42)$$

Similarly to what has been done in the homogeneous case, we will write (4.40) in the compact vector notation

$$D_{|S_i|} \frac{dU}{dt} = -\mathcal{C}U + \mathcal{C}^{\mathcal{S}}\mathcal{S} ,$$

with \mathcal{C} given by (4.29) and $\mathcal{C}^{\mathcal{S}}$ given by

$$c_{ij}^{\mathcal{S}} = \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} |E| c_{ij}^{\mathcal{S}} = |S_i| \tilde{c}_{ij}^{\mathcal{S}}, \quad \tilde{c}_{ij}^{\mathcal{S}} = \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} \frac{|E|}{|S_i|} c_{ij}^{\mathcal{S}} \quad (4.43)$$

4.3.1 L^∞ stability: inhomogeneous case

Uniform bounds on the numerical solution can be derived as in the homogeneous case with respect to one space-time slab. In particular, we will show some simple results relative to the fully discrete analog of (4.40) obtained with the θ -scheme. Using the \tilde{c}_{ij}^S defined in (4.43), the scheme reads

$$|S_i|(u_i^{n+1} - u_i^n) = -\Delta t \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (\theta(u_i^{n+1} - u_j^{n+1}) + (1 - \theta)(u_i^n - u_j^n)) + \Delta t \sum_{j \in \mathcal{D}_i} |S_i| \tilde{c}_{ij}^S S_j \quad (4.44)$$

After introducing the quantities \mathcal{S}_{\min}^i and \mathcal{S}_{\max}^i defined by

$$\mathcal{S}_{\min}^i = \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^S \min_{j \in \mathcal{D}_i} S_j, \quad \mathcal{S}_{\max}^i = \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^S \max_{j \in \mathcal{D}_i} S_j$$

and \mathcal{S}_{\min} and \mathcal{S}_{\max} defined by

$$\mathcal{S}_{\min} = \min_{i \in \mathcal{T}_h} \mathcal{S}_{\min}^i, \quad \mathcal{S}_{\max} = \max_{i \in \mathcal{T}_h} \mathcal{S}_{\max}^i$$

we prove that

Proposition 4.3.1 (L^∞ bounds inhomogeneous case - θ -scheme). *The space-time discrete analog of (4.39) in the time-slab $[t^n, t^{n+1}]$ given by the θ -scheme (4.20) respects the bounds*

$$u_{\min}^n + \Delta t \mathcal{S}_{\min} \leq u_i^{n+1} \leq u_{\max}^n + \Delta t \mathcal{S}_{\max}, \quad (4.45)$$

with u_{\min}^n and u_{\max}^n as in (4.22), if the scheme is LED, if $\forall E \in \mathcal{T}_h$ one has

$$0 \leq \tilde{c}_{ij}^S < \infty \quad \forall i, j \in E$$

and under the time-step constraint (4.24). The explicit FE scheme obtained for $\theta = 0$, also respects the local bounds:

$$\tilde{u}_i^n + \Delta t \mathcal{S}_{\min}^i \leq u_i^{n+1} \leq \tilde{U}_i^n + \Delta t \mathcal{S}_{\max}^i, \quad (4.46)$$

with \tilde{u}_i^n and \tilde{U}_i^n as in proposition 4.1.3.

Proof. As in the homogeneous case, the scheme can be rewritten in the vector form

$$\mathcal{A}U^{n+1} = \mathcal{B}U^n + \Delta t \mathcal{C}^S S$$

with \mathcal{A} and \mathcal{B} as in (4.21) and \mathcal{C}^S as in (4.43). Due to the LED condition, to the time-step constraint and to the positivity of the bounded \tilde{c}_{ij}^S coefficients, both \mathcal{B} and

\mathcal{C}^S contain positive entries. Using (4.21) and the definitions of \mathcal{S}_{\min} and of u_{\min}^n , we obtain the estimates

$$\begin{aligned} (\mathcal{B}U^n + \Delta t \mathcal{C}^S \mathcal{S})_i &= \sum_{j \in \mathcal{D}_i} \mathcal{B}_{ij} u_j^n + \Delta t \sum_{j \in \mathcal{D}_i} |S_i| \tilde{c}_{ij}^S \mathcal{S}_j \geq \\ &\left(\sum_{j \in \mathcal{D}_i} \mathcal{B}_{ij} \right) u_{\min}^n + \Delta t \sum_{j \in \mathcal{D}_i} |S_i| \tilde{c}_{ij}^S \sup_{j \in \mathcal{D}_i} \mathcal{S}_j = \\ &|S_i| (u_{\min}^n + \Delta t \mathcal{S}_{\min}^i) \geq |S_i| (u_{\min}^n + \Delta t \mathcal{S}_{\min}) \end{aligned}$$

Denoting by U_{\min} the array with entries all equal to $u_{\min}^n + \Delta t \mathcal{S}_{\min}$ and proceeding as in the proof of proposition 4.1.7, we note that

$$(\mathcal{B}U^n + \Delta t \mathcal{C}^S \mathcal{S})_i \geq |S_i| (u_{\min}^n + \Delta t \mathcal{S}_{\min}) = (\mathcal{B}U_{\min})_i = (\mathcal{A}U_{\min})_i$$

due to the form of the entries of the \mathcal{A} matrix. As a consequence, we have shown that $(\mathcal{A}U^{n+1})_i \geq (\mathcal{A}U_{\min})_i \forall i \in \mathcal{T}_h$. Using the fact that the LED condition guarantees that \mathcal{A}^{-1} only contains positive entries (see proof of proposition 4.1.7), we get the left inequality in (4.45). The right inequality is obtained using the definition of u_{\max}^n and \mathcal{S}_{\max} and proceeding in a similar way. In the explicit case, \mathcal{A} is diagonal and we can derive the sharper bound:

$$\begin{aligned} u_i^{n+1} &= \left(1 - \frac{\Delta t}{|S_i|} \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij}\right) u_i^n + \sum_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \tilde{c}_{ij} u_j^n + \Delta t \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^S \mathcal{S}_j \geq \\ &\tilde{u}_i^n + \Delta t \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^S \min_{j \in \mathcal{D}_i} \mathcal{S}_j = \tilde{u}_i^n + \Delta t \mathcal{S}_{\min}^i \end{aligned}$$

The right inequality in (4.46) is obtained in a similar way. \square

Clearly, we also have that

Proposition 4.3.2. *The space-time discrete analog of (4.39) in the time-slab $[t^n, t^{n+1}]$ given by the θ -scheme (4.20) respects the bounds (4.45), and in the explicit case the local bounds (4.46), if the sub-element LED condition is satisfied, under the time-step constraint (4.27), and if*

$$c_{ij}^S \geq 0 \quad \forall i, j \in E \quad \text{and} \quad \forall E \in \mathcal{T}_h$$

Application of proposition 4.3.1 over several space-time slabs yields:

Theorem 4.3.3 (L^∞ -stability - inhomogeneous case). *If the hypotheses of proposition 4.3.1 hold in all time slabs $[t^n, t^{n+1}]$, $n = 0, \dots, M-1$ then $\forall t^n = n\Delta t$ scheme (4.44) verifies the L^∞ stability bounds*

$$\min_{j \in \mathcal{T}_h} u_j^0 + t^n \mathcal{S}_{\min} \leq u_i^n \leq \max_{j \in \mathcal{T}_h} u_j^0 + t^n \mathcal{S}_{\max} \quad \forall i \in \mathcal{T}_h$$

Note that, also in the non-homogeneous case, the stability bounds are independent on the mesh size h . In fact, the regularity of the mesh (3.1) implies that we can find positive constants c_1 and c_2 , bounded uniformly with respect to h , such that

$$c_1 \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} c_{ij}^{\mathcal{S}} \min_{j \in \mathcal{D}_i} \mathcal{S}_j \leq \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} \frac{|E|}{|S_i|} c_{ij}^{\mathcal{S}} \min_{j \in \mathcal{D}_i} \mathcal{S}_j = \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^{\mathcal{S}} \min_{j \in \mathcal{D}_i} \mathcal{S}_j = \mathcal{S}_{\min}^i$$

and that

$$c_2 \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} c_{ij}^{\mathcal{S}} \max_{j \in \mathcal{D}_i} \mathcal{S}_j \geq \sum_{E \in \mathcal{D}_i \cap \mathcal{D}_j} \frac{|E|}{|S_i|} c_{ij}^{\mathcal{S}} \max_{j \in \mathcal{D}_i} \mathcal{S}_j = \sum_{j \in \mathcal{D}_i} \tilde{c}_{ij}^{\mathcal{S}} \max_{j \in \mathcal{D}_i} \mathcal{S}_j = \mathcal{S}_{\max}^i$$

Lastly, we add that the analysis can be easily extended to the case $\mathcal{S} = \mathcal{S}(x, y, t)$.

4.4 Accuracy and Godunov's theorem

In the previous sections conditions have been given for the consistency and the stability of the prototype scheme (4.1). Here, we will briefly consider the issue of the accuracy of the approximation obtained by such a scheme. In particular, we will recall conditions under which the approximation is second-order accurate at steady-state and show that, as written in (4.1), the scheme does not allow, in general, to obtain second-order of accuracy in time dependent computations. A more general formulation will have to be considered later to achieve this accuracy. The analysis follows largely [9, 3, 12] and is clearly tailored for the \mathcal{RD} schemes we will present in the next chapter. However, the hypotheses we will do are also valid for \mathcal{FE} schemes, and for \mathcal{FV} schemes, if properly recast according to the abstract representation (4.1).

4.4.1 The steady case

To characterize the accuracy of the approximation, we will consider how well the scheme reproduces the weak formulation of the problem in correspondence of a smooth solution [9, 3, 12]. In particular, we start considering scheme (4.1) at steady-state

$$\sum_{E \in \mathcal{D}_i} \phi_i^E = 0 \quad \forall i \in \mathcal{T}_h$$

and analyze the expression

$$\sum_{i \in \mathcal{T}_h} \varphi_i \left(\sum_{E \in \mathcal{D}_i} \phi_i^E \right) = 0$$

where φ is a smooth compactly supported function $\varphi \in C_0^1(\Omega)$, and $\varphi_i = \varphi(x_i, y_i)$. We denote by φ_h the continuous piecewise linear approximation of φ obtained with the basis functions (3.6)

$$\varphi_h = \sum_{i \in \mathcal{T}_h} \psi_i \varphi_i$$

We suppose now that for our numerical representation of the unknown, $\nabla \cdot \mathcal{F}_h$ is well defined on each $E \in \mathcal{T}_h$, as it is the case *e.g.* for $\mathcal{F} = \bar{a} u_h$ with u_h the continuous piecewise linear interpolation of the nodal values, and that \mathcal{F}_h is continuous across triangle edges, as in the Lax-Wendroff theorem. Introducing in each $E \in \mathcal{T}_h$ the quantity $\bar{\varphi}$, given by the arithmetic average of the nodal values of φ_h in E , using the local consistency assumption and the linearity of φ^h , we can write

$$\begin{aligned}
 0 &= \sum_{i \in \mathcal{T}_h} \varphi_i \left(\sum_{E \in \mathcal{D}_i} \phi_i^E \right) = \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \varphi_i \phi_i^E = \\
 &\quad \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \bar{\varphi} \phi_i^E + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \\
 &\quad \sum_{E \in \mathcal{T}_h} \bar{\varphi} \int_E \nabla \cdot \mathcal{F}_h \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \\
 &\quad \sum_{E \in \mathcal{T}_h} \int_E \varphi_h \nabla \cdot \mathcal{F}_h \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \\
 &\quad \int_{\Omega} \varphi_h \nabla \cdot \mathcal{F}_h \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E
 \end{aligned}$$

Last equation represents the modified equation of our schemes. To see this more clearly we note that for a smooth classical solution of the problem, we can certainly rewrite the last expression as

$$\int_{\Omega} \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = 0$$

which expresses the error between the *approximate* weak formulation of the problem and the analytical one. Proceeding as in [9, 3], we note that for a second-order accurate flux approximation \mathcal{F}_h , and due to the uniform boundedness of $\nabla \varphi_h$, we have

$$\int_{\Omega} \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy = - \int_{\Omega} \nabla \varphi_h \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy = \mathcal{O}(h^2)$$

Hence, if the flux approximation \mathcal{F}_h is second-order accurate, a condition for the method to be second-order accurate is given by

$$\sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} (\varphi_i - \bar{\varphi}) \phi_i^E = \mathcal{O}(h^2)$$

Since the number of nodes in an element is bounded while the number of nodes in a bounded domain is of $\mathcal{O}(h^{-2})$, second-order of accuracy requires that

$$(\varphi_i - \bar{\varphi}) \phi_i^E = \mathcal{O}(h^4)$$

in correspondence of regions containing a smooth solution. Moreover, since $\varphi \in C_0^1$

$$|\varphi_i - \bar{\varphi}| \leq \|\nabla \varphi\|_{L^\infty(\Omega)} h = \mathcal{O}(h) \tag{4.47}$$

As a consequence we are led to [9, 3]

Proposition 4.4.1. *A scheme of the form (4.1) verifying the local consistency (4.3) for a continuous second-order accurate approximation of the flux \mathcal{F}_h , is second-order accurate at steady-state if*

$$\phi_i^E = \mathcal{O}(h^3) \quad (4.48)$$

Strictly speaking, the developments of this section are only exact for constant advection. However, proceeding in a similar way in the general case, on a smooth solution one easily gets to

$$\begin{aligned} 0 &= \sum_{E \in \mathcal{T}_h} \bar{\varphi} \int_E \nabla \cdot \mathcal{F}_h \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \\ &\quad \sum_{E \in \mathcal{T}_h} \bar{\varphi} \int_E \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \\ &\quad \sum_{E \in \mathcal{T}_h} \int_E \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E + \\ &\quad \sum_{E \in \mathcal{T}_h} \int_E (\bar{\varphi} - \varphi_h) \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy \end{aligned}$$

The regularity of φ (equation (4.47)), and the fact that \mathcal{F}_h is a second-order accurate approximation to \mathcal{F} , lead to

$$\int_E (\bar{\varphi} - \varphi_h) \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy = \mathcal{O}(h) \mathcal{O}\left(\oint_{\partial E} (\mathcal{F}_h - \mathcal{F}) \cdot \hat{n} \, dl\right) = \mathcal{O}(h^4)$$

Since the number of elements in \mathcal{T}_h is of order $\mathcal{O}(h^{-2})$, one ends up with

$$\sum_{E \in \mathcal{T}_h} \int_E \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \mathcal{O}(h^2)$$

which ultimately leads to the same result of proposition 4.4.1.

4.4.2 Steady inhomogeneous case

In the case $\mathcal{S}(x, y) \neq 0$, we proceed as before. In particular, for a smooth classical solution, we can write for scheme (4.40)

$$\sum_{E \in \mathcal{T}_h} \int_E \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy - \sum_{E \in \mathcal{T}_h} \int_E \bar{\varphi} (\mathcal{S}_h - \mathcal{S}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = 0$$

We note now that, due to the uniform boundedness of φ_h and of $\nabla \varphi_h$, one has

$$\bar{\varphi} - \varphi_h = \sum_{j \in E} \psi_j (\bar{\varphi} - \varphi_j) = \mathcal{O}(h)$$

Hence, for a second-order accurate approximation of the source term \mathcal{S}_h , and considering that the number of elements in a bounded domain is of $\mathcal{O}(h^{-2})$, we can write

$$\sum_{E \in \mathcal{T}_h} \int_E \bar{\varphi}(\mathcal{S}_h - \mathcal{S}) \, dx \, dy = \sum_{E \in \mathcal{T}_h} \int_E \varphi_h(\mathcal{S}_h - \mathcal{S}) \, dx \, dy + \mathcal{O}(h^3)$$

So that, up to $\mathcal{O}(h^3)$, we obtain the modified equation

$$\int_{\Omega} \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy - \int_{\Omega} \varphi_h(\mathcal{S}_h - \mathcal{S}) \, dx \, dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = 0$$

As in the homogeneous case, if \mathcal{F}_h and \mathcal{S}_h are second-order accurate then

$$\begin{aligned} \int_{\Omega} \varphi_h \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy - \int_{\Omega} \varphi_h(\mathcal{S}_h - \mathcal{S}) \, dx \, dy = \\ - \int_{\Omega} \nabla \varphi_h \cdot (\mathcal{F}_h - \mathcal{F}) \, dx \, dy - \int_{\Omega} \varphi_h(\mathcal{S}_h - \mathcal{S}) \, dx \, dy = \mathcal{O}(h^2) \end{aligned}$$

which is more than the $\mathcal{O}(h^3)$ approximation of the modified equation and shows that, as before, the scheme will be second-order accurate if

$$\sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \phi_i^E = \mathcal{O}(h^2)$$

The same arguments used in the homogeneous case lead to

Proposition 4.4.2. *A scheme of the form (4.40) verifying the local consistency (4.42) for a continuous second-order accurate approximation of the flux \mathcal{F}_h and of the source term \mathcal{S}_h , is second-order accurate at steady-state if*

$$\phi_i^E = \mathcal{O}(h^3) \quad (4.49)$$

4.4.3 Time-dependent computations

The analysis of the inhomogeneous case allows to make a small digression concerning the time-dependent case. The analysis of this case, reported for example in [118, 8], is formally identical to the one of the steady homogeneous case and will be briefly considered in a later chapter. Here, we want to show that, as written in (4.1), the scheme cannot be second-order accurate during transients, due to a lack of *spatial accuracy*. In order to do this, we proceed as in the inhomogeneous case and write:

$$\sum_{i \in \mathcal{T}_h} \varphi_i |S_i| \frac{du_i}{dt} + \sum_{i \in \mathcal{T}_h} \varphi_i \left(\sum_{E \in \mathcal{D}_i} \phi_i^E \right) = 0$$

As before, we manipulate this expression to get

$$\begin{aligned}
 0 &= \sum_{i \in \mathcal{T}_h} \varphi_i |S_i| \frac{du_i}{dt} + \sum_{i \in \mathcal{T}_h} \varphi_i \left(\sum_{E \in \mathcal{D}_i} \phi_i^E \right) = \\
 &= \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} \varphi_i \frac{|E|}{3} \frac{du_i}{dt} + \sum_{i \in \mathcal{T}_h} \varphi_i \left(\sum_{E \in \mathcal{D}_i} \phi_i^E \right) = \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \varphi_i \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right) = \\
 &= \sum_{E \in \mathcal{T}_h} \bar{\varphi} \sum_{i \in E} \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right) + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right) = \\
 &= \sum_{E \in \mathcal{T}_h} \int_E \bar{\varphi} \left(\frac{\partial u_h}{\partial t} + \nabla \cdot \mathcal{F}_h \right) dx dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right)
 \end{aligned}$$

with u_h the piecewise linear continuous interpolation of the nodal values u_i (see equation (3.7)). If we assume that for a smooth classical solution of the problem u , given the second-order accurate approximation u_h , we also have¹

$$\frac{\partial(u_h - u)}{\partial t} = \mathcal{O}(h^2)$$

then, as in the inhomogeneous case, we get (up to $\mathcal{O}(h^3)$)

$$\int_{\Omega} \varphi_h \left(\frac{\partial(u_h - u)}{\partial t} + \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \right) dx dy + \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right) = 0$$

As before, this modified equation gives a condition for the second-order of accuracy of the scheme. In particular, noting that

$$\begin{aligned}
 \int_{\Omega} \varphi_h \left(\frac{\partial(u_h - u)}{\partial t} + \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \right) dx dy &= \\
 \int_{\Omega} \varphi_h \frac{\partial(u_h - u)}{\partial t} dx dy - \int_{\Omega} \nabla \varphi_h \cdot (\mathcal{F}_h - \mathcal{F}) dx dy &= \mathcal{O}(h^2)
 \end{aligned}$$

then second-order of accuracy will be obtained if

$$\sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i - \bar{\varphi}) \left(\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E \right) = \mathcal{O}(h^2)$$

which, once more, leads to the condition

$$\frac{|E|}{3} \frac{du_i}{dt} + \phi_i^E = \mathcal{O}(h^3).$$

However, unless the solution is changing very slowly in time, the first term in last expression will be always of $\mathcal{O}(h^2)$, hence

¹it would perhaps be more correct to assume $\partial(u - u_h)/\partial t = \mathcal{O}(h^2)f(t)$, however we keep the $\mathcal{O}(h^2)$ to simplify the presentation

Proposition 4.4.3. *Given an unstructured discretization of the spatial domain Ω , a scheme of the form (4.1) verifying the local consistency (4.3) for a continuous second-order accurate approximation of the flux \mathcal{F}_h , is first-order accurate in space in time-dependent computations.*

As it was mentioned in the introduction, there have been studies in literature aiming at constructing schemes which apparently are exceptions to last proposition. With the exception of one result published very recently in [62], these Lax-Wendroff type of schemes have only shown second-order of accuracy on structured triangulations [90, 60]. As shown in [143] on structured grids a fortunate error cancellation occurs, leading to a second-order discretization. As we will show later, a consistent construction leads to the introduction of a mass matrix also for these schemes. Even though the analysis of this section is not very rigorous, its final output has indeed a general (and well known) character. Note also that the analysis only takes into account the accuracy in *in space*: scheme (4.1) lacks spatial accuracy in the time-dependent case. Hence, high-order time integration cannot cure this problem. Different forms of schemes will have to be considered to be able to retain second-order of accuracy in unsteady computations.

4.4.4 Linear schemes and Godunov's theorem

We end this chapter by recalling a result which is a generalization of Godunov's theorem [77] to the scheme analyzed in this chapter. First, we give the following definition

Definition 4.4.4 (Linear scheme). *A scheme of the form (4.1) is said to be linear if all the c_{ij}^E are independent on the numerical solution.*

Finally we have the following theorem [126, 9, 118].

Theorem 4.4.5. *No linear scheme of the form (4.1) can be simultaneously positive and second-order accurate.*

4.5 Summary

We have introduced a prototype compact scheme for scalar advection and analyzed under which conditions the scheme respects discrete analogs of the stability properties of exact solutions. Conditions for achieving second-order of accuracy have been also shown. The main results are summarized hereafter.

- The solution exhibits a discrete maximum principle, provided that the coefficients in the discretization verify a *positivity* condition, and under a time-step restriction;
- The time-step restrictions also apply to implicit schemes, with the exception of backward Euler time integration;
- The L^∞ stability has been extended to inhomogeneous problems where the source term does not depend on the solution;
- Positive schemes also respect some form of energy stability. However, when discretizing the time derivative, additional terms appear in the energy balance which can stabilize or destabilize the schemes;
- Second-order accurate schemes for steady inhomogeneous problems can be constructed. A condition ensuring this level of accuracy has been given;
- The prototype scheme considered in this chapter cannot be second-order accurate during transients. This is due to a lack of accuracy in space, hence high-order time integration does not solve the problem;
- Linear schemes can either be L^∞ -stable or second-order accurate due to Godunov's theorem.

Once more we remark that, even though the prototype scheme presented fits perfectly in the residual distribution framework considered in the thesis, a large number of \mathcal{FV} and \mathcal{FE} schemes are included in this abstract model, as we will show in the next chapter. This gives more generality to our analysis.

Chapter 5

$\mathcal{RD}/\mathcal{FS}$ schemes for steady advection

In this chapter we finally introduce the schemes which are at the basis of our work: the Residual Distribution (\mathcal{RD}) or Fluctuation Splitting (\mathcal{FS}) schemes. We start by giving a precise definition of what a \mathcal{RD} scheme is. It will be immediately clear that (4.1) is indeed a prototype for these schemes, although a subtle difference in the role of the consistency condition (4.3) exists. To show this, we give examples of \mathcal{FV} and \mathcal{FE} schemes which, although constructed on completely different grounds, can be recast into the \mathcal{RD} formalism. Finally, we present the multidimensional upwind schemes at the basis of this work. Throughout the presentation, we focus as much as possible on geometrical aspects of the discretization, even though some theoretical issues are covered as well. Illustrative computational examples are given at the end of the chapter to *experimentally visualize* the differences between the schemes presented.

5.1 $\mathcal{RD}/\mathcal{FS}$: definition and generalities

We define here schemes for the numerical solution of the scalar advection equation

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \mathcal{S}(x, y) \quad \text{on } \Omega \times [0, t_f]$$

on an unstructured discretization of the spatial domain Ω . Note that in this chapter, as in the following ones, we will make extensive use of the notation introduced in chapter 2, chapter 3 and chapter 4, to which the reader is referred for clarifications.

Given an initial solution $u_0(x, y)$, in this thesis we are interested in the following class of discretizations

Definition 5.1.1 (Residual Distribution/Fluctuation Splitting scheme). A Residual Distribution or Fluctuation Splitting scheme is defined as a scheme that, given the continuous approximation of the initial solution u_h^0 as in equation (3.8), given the continuous approximation of the unknown u_h as in (3.7), and the continuous approximation of the source term \mathcal{S}_h (3.15), evolves in time the nodal values of u_h as follows

1. $\forall E \in \mathcal{T}_h$ compute the residual or fluctuation

$$\phi^h = \int_E (\vec{a} \cdot \nabla u_h - \mathcal{S}_h) \, dx \, dy = \int_{\partial E} \mathcal{F}_h \cdot \hat{n} \, dl - \int_E \mathcal{S}_h \, dx \, dy \quad (5.1)$$

2. $\forall E \in \mathcal{T}_h$ distribute fractions of ϕ^h to each node of E . Denoting by ϕ_i the split residual or local nodal residual for node $i \in E$, by construction one must have

$$\sum_{j \in E} \phi_j = \phi^h \quad (5.2)$$

Equivalently, denoting by β_i the distribution coefficient of node i :

$$\beta_i = \frac{\phi_i}{\phi^h} \quad (5.3)$$

one must have by construction

$$\sum_{j \in E} \beta_j = 1 \quad (5.4)$$

3. $\forall i \in \mathcal{T}_h$ assemble the elemental contributions from all $E \in \mathcal{D}_i$ and evolve u_i in time according to (see equation (4.1))

$$|S_i| \frac{du_i}{dt} + \sum_{E \in \mathcal{D}_i} \phi_i = 0 \quad (5.5)$$

Note that, to simplify the notation, in the previous definition the superscript E has been removed from the local nodal residual ϕ_i . The element E of the mesh to which this quantity refers is always clear from the context. It is evident that definition 5.1.1 would reduce to the prototype (4.1) unless the continuity of u_h (and more generally of the flux $\mathcal{F}_h = (\vec{a}u)_h$) and the consistency requirement (5.2) were not imposed *by construction*. In other words, for these \mathcal{FS} schemes the prototype (4.1) with the consistency assumption (4.3) verified for a continuous numerical approximation of the flux \mathcal{F}_h (as required by the Lax-Wendroff theorem), is a natural generalization.

We give in the following sections an overview of the properties of \mathcal{RD} schemes. Even though references are given throughout the text, it is best for the interested reader to consult the extensive bibliography given in chapter 1 for a comprehensive historical overview, and for a relatively up-to-date list of papers describing the most recent developments.

5.1.1 The residual

For this simple problem, it is possible to express the fluctuation ϕ^h in a closed form allowing to give a first description of the potential of the \mathcal{RD} approach. We start by noting that, for the numerical approximation of the unknown u_h given in (3.7), in each element $E \in \mathcal{T}_h$ the gradient of u_h is constant and given by (see also equation (3.6))

$$\nabla u_h|_E = \sum_{j \in E} u_j \frac{\vec{n}_j}{2|E|} \quad (5.6)$$

Hence, for a constant advection speed \vec{a} , in the homogeneous case the residual can be expressed using the k_j parameters introduced in chapter 3 (equation (3.27)):

$$\phi^h = \sum_{j \in E} k_j u_j . \quad (5.7)$$

Note that, due to their definition, the k_j parameters can be used as *sensors* to distinguish between down-stream and up-stream nodes. In particular, $k_j > 0$ only if \vec{a} is oriented as \vec{n}_j , hence only if node j is down-stream. For scalar advection, the case $\vec{a} = \vec{a}(x, y)$ is treated similarly, using in the computations of the k_j parameters the average value

$$\bar{a} = \frac{1}{|E|} \int_E \vec{a}(x, y) \, dx \, dy \quad (5.8)$$

For this reason, in the following we shall assume that \vec{a} is constant over E , implicitly assuming that the linearization (5.8) has been used in the general case.

Expression (5.7) can be recast in an interesting alternate form. In particular, using the upwind parameters (3.27) and the identity (3.20), one has

$$\phi^h = \sum_{j \in E} k_j^+ u_j + \sum_{j \in E} k_j^- u_j = \left(\sum_{j \in E} k_j^+ \right) \left(\sum_{j \in E} N k_j^+ u_j + \sum_{j \in E} N k_j^- u_j \right)$$

having introduced the quantity

$$N = \left(\sum_{j \in E} k_j^+ \right)^{-1} = - \left(\sum_{j \in E} k_j^- \right)^{-1} = \frac{1}{2} \left(\sum_{j \in E} |k_j| \right)^{-1} > 0 \quad (5.9)$$

Defining the *inflow* and *outflow states* of element E

$$u_{in} = \frac{\sum_{j \in E} k_j^- u_j}{\sum_{j \in E} k_j^-} = - \sum_{j \in E} N k_j^- u_j \quad \text{and} \quad u_{out} = \sum_{j \in E} N k_j^+ u_j \quad (5.10)$$

the residual can be written as [129]

$$\phi^h = \left(\sum_{j \in E} k_j^+ \right) (u_{out} - u_{in}) . \quad (5.11)$$

First of all, we note that u_{in} and u_{out} are a convex combination of the nodal values $\{u_j\}_{j \in E}$, hence they are bounded by the minimum and maximum value of u_h over E .

Proposition 5.1.2. *The inflow/outflow state defined by equation (5.10) respects*

$$\min_{j \in E} u_j \leq u_{in} \leq \max_{j \in E} u_j \quad (5.12)$$

$$\min_{j \in E} u_j \leq u_{out} \leq \max_{j \in E} u_j \quad (5.13)$$

A neat geometrical interpretation of (5.11) can be given by noting that the inflow and outflow states represent the values of u_h in the most upstream (resp. most downstream) node of the E , with respect to the streamline cutting the triangle [129, 28]:

$$u_{out} = u_h(\vec{x}_{out}), \quad \vec{x}_{out} = \sum_{j \in E} N k_j^+ \vec{x}_j$$

$$u_{in} = u_h(\vec{x}_{in}), \quad \vec{x}_{in} = - \sum_{j \in E} N k_j^- \vec{x}_j$$

As a consequence, the residual (5.11) represents a onedimensional balance along ζ , the streamline cutting the element. Clearly, this framework gives the basis, at least in principle, for a truly multidimensional generalization of concepts derived from the study of onedimensional advection. In particular, as depicted in figure 5.1, we can distinguish two situations, depending on how \vec{a} is oriented in E .

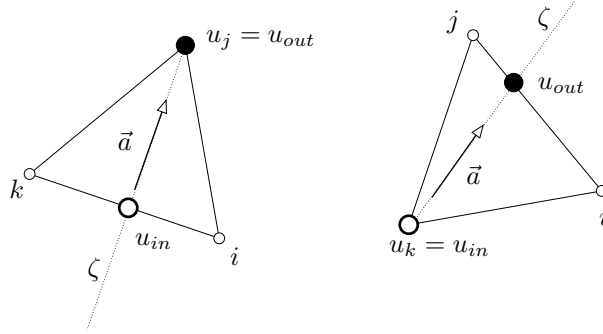


Figure 5.1: Inflow and outflow state. One-target (left) and two-target element (right)

If \vec{a} points in the direction of a single point of E , as in the left picture on the figure, then this point coincides with the outflow point and is the only downstream point. In this situation the element is said to be a *one-target* element. Conversely, if \vec{a} points in the direction of one of the edges of E , as in the right picture, then there is only one upstream point coinciding with the inflow point. In this situation the element is said to be a *two-target* element. If E is one-target, then there is a node j such that

$$k_j = k_j^+ > 0, \quad k_j^- = 0 \quad \text{and} \quad k_l = k_l^- < 0, \quad k_l^+ = 0 \quad \forall l \neq j$$

Similarly, if E is two-target, then there is a node k such that

$$k_k = k_k^- < 0, \quad k_k^+ = 0 \quad \text{and} \quad k_l = k_l^+ > 0, \quad k_l^- = 0 \quad \forall l \neq k$$

This distinction allows to build discretizations taking into account in a real multidimensional way the propagation of the information described by the advection equation.

In the inhomogeneous case, using (3.15) one easily shows that

$$\phi^h = \sum_{j \in T} k_j u_j - \sum_{j \in T} \frac{|E|}{3} \mathcal{S}_j. \quad (5.14)$$

It might seem useful to manipulate this expression, using the information on the direction of the advection speed, in order to distinguish between upstream nodes and downstream nodes also for the source term. In practice this complication has not given way to successful schemes [173], or not yet at least.

5.1.2 A residual property: linearity preserving schemes

Consider a steady homogeneous problem admitting a smooth exact solution u . If the advection speed verifies $\nabla \cdot \vec{a} = 0$, the definition of the residual gives

$$\begin{aligned} \phi^h &= \int_E \vec{a} \cdot \nabla u_h \, dx \, dy = \int_E \nabla \cdot (\vec{a} u_h) \, dx \, dy = \\ &= \int_E \nabla \cdot (\vec{a} u_h - \vec{a} u) \, dx \, dy = \oint_{\partial E} (u_h - u) \vec{a} \cdot \hat{n} \, dl = \mathcal{O}(h^3) \end{aligned}$$

since u is a smooth solution, $|\partial E| = \mathcal{O}(h)$ and \vec{a} is bounded. Similarly, for a non-homogeneous problem admitting a smooth steady solution, the element residual reads

$$\begin{aligned} \phi^h &= \int_E (\vec{a} \cdot \nabla u_h - \mathcal{S}_h) \, dx \, dy = \int_E (\nabla \cdot (\vec{a} u_h) - \mathcal{S}_h) \, dx \, dy = \\ &= \int_E (\nabla \cdot (\vec{a} u_h - \vec{a} u) - (\mathcal{S}_h - \mathcal{S})) \, dx \, dy = \\ &= \oint_{\partial E} (u_h - u) \vec{a} \cdot \hat{n} \, dl - \int_E (\mathcal{S}_h - \mathcal{S}) \, dx \, dy = \mathcal{O}(h^3) + \mathcal{O}(h^4) = \mathcal{O}(h^3) \end{aligned}$$

These estimates, combined with propositions 4.4.1 and 4.4.2, lead to the result that if the distribution coefficients (5.3) are uniformly bounded with respect to the solution and the data of the problem, then one has

$$\phi_i = \beta_i \phi^h = \mathcal{O}(h^3)$$

hence the scheme is second-order accurate at steady-state [3, 9, 12]:

Definition 5.1.3 (Linearity Preserving scheme). A \mathcal{RD} scheme is linearity preserving (\mathcal{LP}) if its distribution coefficients are uniformly bounded with respect to the solution and the data of the problem:

$$\max_{E \in \mathcal{T}_h} \max_{j \in E} |\beta_j| < C < \infty \quad \forall \phi^h, u_h, \vec{a}, u_h^0, \dots$$

Linearity preserving schemes are second-order accurate at steady-state by construction.

As shown in chapter 4, second-order of accuracy is obtained *only* at steady-state. However, the idea of a distribution of the residual with bounded coefficients is at the basis of the construction of all high-order \mathcal{RD} schemes, since it allows to build accurate discretizations just by properly defining the fluctuation ϕ^h . We recall that, due to Godunov's theorem, these schemes cannot be also positive, unless some nonlinearity is introduced, that is, some dependence of the distribution coefficients on the solution.

5.2 Finite Volume schemes in \mathcal{FS} formalism

After having precisely defined a \mathcal{RD} scheme, we show that some discretization techniques, constructed on completely different grounds, can be recast into a \mathcal{FS} formalism, justifying the introduction of the abstract model (4.1). In particular, we recall in this section a well known equivalence between first-order finite volume schemes on the median dual cell [17, 19, 20, 23, 24, 25] and \mathcal{RD} . The analysis follows [5, 6].

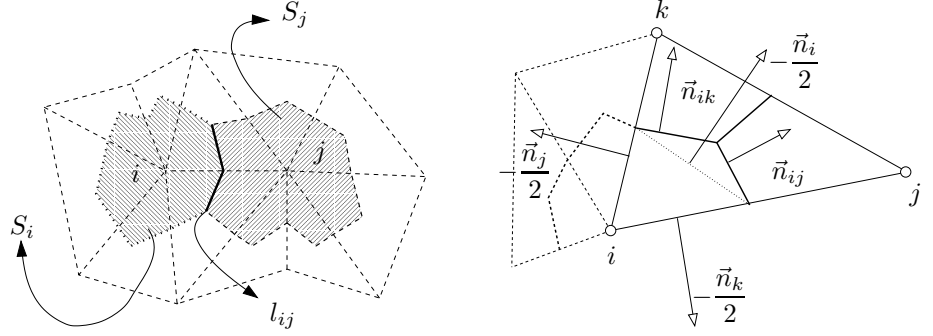


Figure 5.2: \mathcal{FV} scheme. Neighboring cells S_i and S_j (left) and cell normals (right)

Consider then a piecewise constant numerical approximation $u'_h \in \mathcal{S}_h$, with

$$\mathcal{S}_h = \{u'_h; u'_h|_{S_i} \text{ is constant } \forall i \in \mathcal{T}_h\} \quad (5.15)$$

The \mathcal{FV} semi-discrete counterpart of the scalar advection equation reads

$$|S_i| \frac{du_i}{dt} = - \oint_{\partial S_i} \mathcal{F}_h(u'_h) \cdot \vec{n} \, dl = - \sum_{l_{ij} \in \partial S_i} \int_{l_{ij}} H(u_i, u_j) \cdot \vec{n}_{ij} \, dl$$

where $H(u, v)$ is the \mathcal{FV} numerical flux, l_{ij} is the portion of ∂S_i separating S_i from S_j (see left picture on figure 5.2), and \vec{n}_{ij} is the exterior unit normal to ∂S_i on l_{ij} . We are interested in first-order schemes for which last expression becomes

$$|S_i| \frac{du_i}{dt} = - \sum_{l_{ij} \in \partial S_i} H_h(u_i, u_j) \cdot \vec{n}_{ij}$$

with the scaled exterior normal $\vec{n}_{ij} = |l_{ij}| \vec{n}_{ij}$, as in the right picture on figure 5.2. With reference to this picture, we can easily recast the right hand side in last equation

as a sum of contributions coming from elements in \mathcal{D}_i :

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} H(u_i, u_j) \cdot \vec{n}_{ij}$$

This already shows that the scheme can be written with the abstract formalism of (4.1). Moreover, thanks to the consistency property of the \mathcal{FV} flux

$$H(u, u) = \mathcal{F}(u) = \vec{a}u$$

we can add, in each element E , flux contributions coming from the portions of the edges of E contained in S_i , since they cancel identically when summing over all the elements in \mathcal{D}_i (see right picture on figure 5.2):

$$\sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} H(u_i, u_i) \cdot \frac{\vec{n}_j}{2} = \mathcal{F}(u_i) \cdot \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} \frac{\vec{n}_j}{2} = \mathcal{F}(u_i) \cdot \sum_{E \in \mathcal{D}_i} \frac{\vec{n}_i}{2} = 0$$

hence

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} \left(H(u_i, u_j) \cdot \vec{n}_{ij} - H(u_i, u_i) \cdot \frac{\vec{n}_j}{2} \right)$$

Moreover, due to the definition of the median dual cell, we also have (see figure 5.2)

$$\sum_{\substack{j \in E \\ j \neq i}} \frac{\vec{n}_j}{2} = - \frac{\vec{n}_i}{2} = \sum_{\substack{j \in E \\ j \neq i}} \vec{n}_{ij} \quad (5.16)$$

leading to

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} (H(u_i, u_j) - H(u_i, u_i)) \cdot \vec{n}_{ij}$$

We consider now the family of flux functions defined as

$$H(u_i, u_j) = \frac{\mathcal{F}(u_i) + \mathcal{F}(u_j)}{2} \cdot \vec{n}_{ij} - \frac{1}{2} D(u_i, u_j)(u_j - u_i) \quad (5.17)$$

with $D(u_i, u_j)$ a *dissipation matrix* (e.g. Roe's absolute value matrix [147]) satisfying the symmetry condition $D(u_i, u_j) = D(u_j, u_i)$. With this definition the \mathcal{FV} scheme can be written as

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \phi_i = - \frac{1}{2} \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} ((\mathcal{F}(u_j) - \mathcal{F}(u_i)) \cdot \vec{n}_{ij} - D(u_i, u_j)(u_j - u_i)) \quad (5.18)$$

In order for last expression to define a \mathcal{FS} scheme, the ϕ_i s must verify the consistency condition (5.2), for a *continuous* approximation of the flux \mathcal{F}_h . Hence, we compute

$$\begin{aligned} \sum_{i \in E} \phi_i &= \sum_{i \in E} \frac{1}{2} \sum_{\substack{j \in E \\ j \neq i}} ((\mathcal{F}(u_j) - \mathcal{F}(u_i)) \cdot \vec{n}_{ij} - D(u_i, u_j)(u_j - u_i)) = \\ &= \sum_{i \in E} \frac{1}{2} \sum_{\substack{j \in E \\ j \neq i}} (\mathcal{F}(u_j) - \mathcal{F}(u_i)) \cdot \vec{n}_{ij} \end{aligned}$$

since the dissipation terms cancel mutually when summing over the nodes, thanks to the symmetry of $D(u_i, u_j)$. Finally, using the relation $\vec{n}_{ij} = -\vec{n}_{ji}$ and (5.16), one easily shows that

$$\sum_{i \in E} \phi_i = \sum_{i \in E} \frac{1}{2} \mathcal{F}(u_i) \cdot \vec{n}_i$$

which, in the homogeneous case, corresponds to (5.1) integrated exactly for a *continuous piecewise linear approximation of the flux* \mathcal{F}_h . This shows that any finite volume scheme operating on the median dual cells with numerical flux function of the type (5.17) is equivalent to the \mathcal{RD} scheme with the local nodal residuals (implicitly) defined in (5.18), obtained with a continuous linear approximation of the flux. Note that the analysis is general and can be extended to nonlinear problems and systems. Moreover, as shown in [5, 6], it applies to general \mathcal{FV} numerical fluxes and not only to (5.18). Surprisingly, starting from the piecewise constant \mathcal{FV} approximation we have arrived to a scheme based on a continuous flux approximation which, moreover, respects all the hypotheses of the Lax-Wendroff theorem for the prototype scheme (4.1) (see chapter 4 and [5, 6]). Concerning this theorem, we also mention the early work of [133].

5.2.1 The upwind \mathcal{FV} scheme: positivity and energy stability

For scalar advection, the most natural choice for $H(u, v)$, is the *upwind flux*

$$H(u_i, u_j) = \frac{\mathcal{F}(u_i) + \mathcal{F}(u_j)}{2} \cdot \vec{n}_{ij} - \frac{1}{2} \left| \frac{\partial \mathcal{F}}{\partial u} \cdot \vec{n}_{ij} \right|_{ij} (u_j - u_i)$$

which for this linear problem reduces to

$$H(u_i, u_j) = k_{ij} \frac{(u_i + u_j)}{2} - \frac{|k_{ij}|}{2} (u_j - u_i), \quad k_{ij} = \vec{a} \cdot \vec{n}_{ij} \quad (5.19)$$

Hence, the \mathcal{FV} semi-discrete equation becomes

$$|S_i| \frac{du_i}{dt} = - \sum_{E \in \mathcal{D}_i} \phi_i = - \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} \left(k_{ij} \frac{(u_j - u_i)}{2} - \frac{|k_{ij}|}{2} (u_j - u_i) \right)$$

which finally leads to the upwind $\mathcal{FV} - \mathcal{RD}$ scheme defined by the split residuals [129]

$$\phi_i^{\mathcal{FV} - \mathcal{RD}} = - \sum_{\substack{j \in E \\ j \neq i}} k_{ij}^- (u_i - u_j) \quad (5.20)$$

Scheme (5.20) is of the form (4.1) with $c_{ij}^E = -k_{ij}^- \geq 0$, hence it respects the sub-element LED condition, hence it verifies propositions 4.1.3, 4.1.4, 4.1.7, 4.1.8 and 4.1.5, and theorems 4.1.6 and 4.1.9, and the related stability bounds. In particular, the time-step restriction for its positivity reads

$$\Delta t \leq \frac{|S_i|}{(1 - \theta) \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} -k_{ij}^-} \quad \theta \in [0, 1). \quad (5.21)$$

Similarly, its local positivity is constrained by

$$\Delta t \leq \frac{|E|}{3(1-\theta) \sum_{\substack{j \in E \\ j \neq i}} -k_{ij}^-} \quad \theta \in [0, 1), \quad \forall i \in E \quad \forall E \in \mathcal{T}_h. \quad (5.22)$$

The upwind $\mathcal{FV} - \mathcal{RD}$ scheme is unconditionally positive when backward Euler time-integration is used in (5.5). For this scheme the distribution coefficients are not explicitly defined. Instead, they have to be computed as

$$\beta_i^{\mathcal{FV}-\mathcal{RD}} = \frac{\phi_i^{\mathcal{FV}-\mathcal{RD}}}{\phi^h}$$

Since $\beta_i^{\mathcal{FV}-\mathcal{RD}}$ is not guaranteed to be bounded as $\phi^h \rightarrow 0$, the scheme is not \mathcal{LP} , in accordance with Godunov's theorem. However, we note that

$$\begin{aligned} \sum_{j \in E} (c_{ij}^E - c_{ji}^E) &= - \sum_{j \in E} (k_{ij}^- - k_{ji}^-) = - \frac{1}{2} \sum_{j \in E} (k_{ij} - |k_{ij}| - k_{ji} + |k_{ji}|) = \\ &= - \frac{1}{2} \sum_{j \in E} (k_{ij} - |k_{ij}| + k_{ij} + |k_{ij}|) = - \sum_{j \in E} k_{ij} = k_i \end{aligned}$$

since $k_{ij} = -k_{ji}$ and making use of (5.16) and of the definitions of k_{ij} and k_i , equations (5.19) and (3.27) respectively.

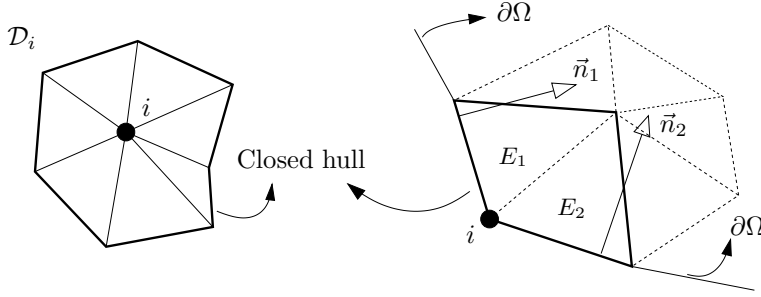


Figure 5.3: Closed hull around node i

Since the hull composed by the edges opposite to i is closed (see left picture on figure 5.3), then for constant scalar advection

$$\sum_{E \in \mathcal{D}_i} k_i = \frac{1}{2} \vec{a} \cdot \sum_{E \in \mathcal{D}_i} \vec{n}_i = 0$$

which proves that the upwind $\mathcal{FV} - \mathcal{RD}$ scheme respects the energy stability criteria of propositions 4.2.2 and 4.2.7 and of proposition 4.2.8 in the fully discrete case. Note that, with reference to the right picture on figure 5.3, for a boundary node $i \in \partial\Omega$ the last sum is not zero but it is given by:

$$\sum_{E \in \mathcal{D}_i} k_i = -\frac{1}{2} \vec{a} \cdot (\vec{n}_1 + \vec{n}_2)$$

with the inward normals \vec{n}_1 and \vec{n}_2 scaled by the length of the edges. In particular, one easily shows that, when included in the energy balance, these terms give a second-order approximation of the energy flux across $\partial\Omega$, so that the global energy estimate becomes (see equation (4.31) and proposition 4.2.7)

$$\frac{d\mathcal{E}_h}{dt} = -U^T \overline{M}^{\mathcal{E}_h} U - \frac{1}{2} \oint_{\partial\Omega} u_h (\vec{a} \cdot \hat{n}) u_h dl$$

with \hat{n} the exterior unit normal to $\partial\Omega$ and $\overline{M}^{\mathcal{E}_h}$ positive semi-definite. How to handle the extra BCs terms will be shown in the next section.

5.3 Central schemes and \mathcal{FE}

In this section we consider a second family of schemes which originally are not formulated as \mathcal{FS} schemes but that fall in the \mathcal{RD} formalism. They are all variations of a central scheme obtained by equi-distributing the residual to the nodes of an element. In particular, we show that among these schemes, the \mathcal{LP} ones are finite element schemes.

5.3.1 The central \mathcal{RD} scheme and the Galerkin \mathcal{FE} method

We consider first the Galerkin \mathcal{FE} scheme. For steady constant advection, and neglecting the BC terms, the scheme reads

$$\int_{\Omega} \psi_i \vec{a} \cdot \nabla u_h dx dy = 0 \quad \forall i \in \mathcal{T}_h \quad (5.23)$$

where the ψ_i s are the linear basis functions introduced in chapter 2, and the approximation of the unknown u_h is as in (3.7). In the case of a constant advection speed, using the compactness of the support of the basis functions and (3.6), the Galerkin scheme can be immediately recast as

$$\sum_{E \in \mathcal{D}_i} \frac{1}{3} \phi^h = 0 \quad \forall i \in \mathcal{T}_h$$

which is nothing else but the steady-state discrete approximation of the advection equation with the \mathcal{LP} fluctuation splitting scheme with distribution coefficients

$$\beta_i^C = \frac{1}{3} \quad (5.24)$$

This *centered* \mathcal{RD} scheme is then exactly equivalent to the \mathcal{FE} Galerkin scheme, if the advection speed \vec{a} is constant.

5.3.2 Petrov-Galerkin schemes in \mathcal{RD} form

The Galerkin method is known to be unstable when approximating the advection equation. Consider then the stabilized Petrov-Galerkin (PG) schemes, obtained by adding to the Galerkin discretization a so-called *streamline dissipation* term [72, 94, 97, 98, 96, 102, 103, 166]:

$$\int_{\Omega} \psi_i \vec{a} \cdot \nabla u_h \, dx \, dy + \underbrace{\sum_{E \in \mathcal{T}_h} \int_E \tau (\vec{a} \cdot \nabla \psi_i) (\vec{a} \cdot \nabla u_h) \, dx \, dy}_{\text{PG streamline dissipation}} = 0 \quad \forall i \in \mathcal{T}_h \quad (5.25)$$

In the case of a constant advection speed \vec{a} , proceeding as before, we quickly arrive to

$$0 = \sum_{E \in \mathcal{D}_i} \frac{1}{3} \phi^h + \sum_{E \in \mathcal{D}_i} \tau \frac{k_i}{2|E|} \phi^h = \sum_{E \in \mathcal{D}_i} \phi_i^C + \sum_{E \in \mathcal{D}_i} \tau \frac{k_i}{2|E|} \phi^h \quad \forall i \in \mathcal{T}_h \quad (5.26)$$

Which shows the equivalence of stabilized Petrov-Galerkin \mathcal{FE} schemes with the class of linearity preserving \mathcal{RD} schemes defined by the distribution coefficients

$$\beta_i^{\text{PG}} = \frac{1}{3} + \tau \frac{k_i}{2|E|} \quad \tau \geq 0 \quad (5.27)$$

Note that the schemes are indeed consistent thanks to relation (3.17). This analogy is of course known for a long time (see for example [127, 39, 176, 129] and references therein). We remark, however, that strictly speaking, the analogy is an equivalence only in the constant coefficients case, while in general the \mathcal{RD} scheme and the \mathcal{FE} schemes give different discrete equations, since the integrals in (5.25) no more reduce to (5.26). Note also that the streamline dissipation terms, introduce some kind of upwind bias in the distribution, since we have

$$\begin{aligned} \beta_i^{\text{PG}} &> \beta_i^C && \text{if } i \text{ is downstream, hence } k_i > 0 \\ \beta_i^{\text{PG}} &< \beta_i^C && \text{if } i \text{ is upstream, hence } k_i < 0 \end{aligned}$$

The stabilization mechanism introduced by this upwind bias is better understood by looking at the energy stability of the schemes. This will be also useful to gain more understanding of the multidimensional upwind schemes we are going to introduce in the next section, and also in view of the nonlinear analysis we will perform in the next chapter.

5.3.2.1 PG schemes: energy stability

Being \mathcal{LP} , the PG- \mathcal{RD} schemes are not LED. However, they have well known energy stability properties that we will recall here. To derive an energy estimate we start by constructing the local energy production (4.36)

$$\frac{d\mathcal{E}^{\text{PG}}}{dt} = \frac{d\mathcal{E}^C}{dt} - \epsilon^{\text{PG}} = - \sum_{i \in E} \sum_{j \in E} \frac{1}{3} u_i k_j u_j - \sum_{i \in E} \sum_{j \in E} u_i \frac{k_i \tau k_j}{2|E|} u_j \quad (5.28)$$

Due to the properties of the basis functions, if \vec{a} is constant, one easily shows that

$$\sum_{i \in E} \sum_{j \in E} \frac{1}{3} u_i k_j u_j = \int_E u_h \vec{a} \cdot \nabla u_h \, dx \, dy$$

Moreover, the second term can be written as

$$\epsilon^{\text{PG}} = \frac{1}{2|E|} \begin{bmatrix} k_1 u_1 \\ k_2 u_2 \\ k_3 u_3 \end{bmatrix}^T \begin{bmatrix} \tau & 0 & 0 \\ 0 & \tau & 0 \\ 0 & 0 & \tau \end{bmatrix} \begin{bmatrix} k_1 u_1 \\ k_2 u_2 \\ k_3 u_3 \end{bmatrix} \geq 0 \quad \text{if } \tau \geq 0 \quad (5.29)$$

which shows that the upwind bias introduced by the streamline diffusion operator adds a stabilizing dissipation mechanism¹. Indeed, assembling the contributions of all the elements in the mesh, the energy balance becomes (see also equation (4.31))

$$\frac{d\mathcal{E}_h^{\text{PG}}}{dt} = - \sum_{E \in \mathcal{T}_h} \int_E u_h \vec{a} \cdot \nabla u_h \, dx \, dy - \sum_{E \in \mathcal{T}_h} \epsilon^{\text{PG}} = - \int_{\Omega} u_h \vec{a} \cdot \nabla u_h \, dx \, dy - \epsilon_h^{\text{PG}} \quad (5.30)$$

with

$$\epsilon_h^{\text{PG}} = \sum_{E \in \mathcal{T}_h} \epsilon^{\text{PG}} \geq 0$$

This only shows that a dissipative mechanism is present, unless the boundary conditions are taken into account. For simplicity, we suppose that homogeneous BCs are prescribed. To be completely faithful to the \mathcal{FE} formulation, the BCs should be included in the variational formulation (5.25) using the admissibility condition [19]

$$\min(\vec{a} \cdot \hat{n}, 0)u = (\vec{a} \cdot \hat{n})^- u = 0 \quad \text{on } \partial\Omega$$

with \hat{n} the unit exterior normal to $\partial\Omega$. However, here we suppose that the BCs are imposed in a strong nodal sense, such that

$$\oint_{\partial\Omega} u_h (\vec{a} \cdot \hat{n})^- u_h \, dl = 0 \quad (5.31)$$

either because we impose $u_h = 0$ or because $(\vec{a} \cdot \hat{n})^- = 0$. We then rewrite the energy estimate (5.30) as

$$\begin{aligned} \frac{d\mathcal{E}_h^{\text{PG}}}{dt} &= - \int_{\Omega} u_h \vec{a} \cdot \nabla u_h \, dx \, dy - \epsilon_h^{\text{PG}} = - \frac{1}{2} \oint_{\partial\Omega} u_h (\vec{a} \cdot \hat{n}) u_h \, dl - \epsilon_h^{\text{PG}} = \\ &\quad - \frac{1}{2} \oint_{\partial\Omega} u_h |\vec{a} \cdot \hat{n}| u_h \, dl - \oint_{\partial\Omega} u_h (\vec{a} \cdot \hat{n})^- u_h \, dl - \epsilon_h^{\text{PG}} \end{aligned}$$

where the identity $\vec{a} \cdot \hat{n} = 2(\vec{a} \cdot \hat{n})^- + |\vec{a} \cdot \hat{n}|$ has been used. Finally, using the strong BCs (5.31), we obtain the energy stability estimate

$$\frac{d\mathcal{E}_h^{\text{PG}}}{dt} = - \frac{1}{2} \oint_{\partial\Omega} u_h |\vec{a} \cdot \hat{n}| u_h \, dl - \epsilon_h^{\text{PG}} \leq 0 \quad (5.32)$$

¹hence the name...

²which we are allowed to do only if $(\vec{a} \cdot \hat{n})^- \neq 0$

As already remarked, a faithful analysis would have included the boundary conditions directly into the variational formulation. This, however, would have led precisely to estimate (5.32) [19]. The most important point of the analysis is that it shows that the total energy production can be split into the energy dissipation introduced by the upwind biasing (ϵ_h^{PG}) plus the energy production due to the centered discretization terms. The latter is then simplified taking into account the boundary conditions, finally obtaining an energy stability estimate. This observation will be useful in the analysis of the stability of the multidimensional upwind schemes.

5.3.3 The Rusanov scheme

Among central \mathcal{RD} schemes, there exists also a LED scheme known as the Rusanov's (Rv) scheme [9, 10], and defined by the split residuals

$$\phi_i^{\text{Rv}} = \frac{1}{3}\phi^h + \frac{1}{3}\alpha \sum_{\substack{j \in E \\ j \neq i}} (u_i - u_j), \quad \alpha \geq \max_{j \in E} |k_j| > 0 \quad (5.33)$$

The Rv scheme is then obtained from the centered scheme by adding to it a stabilizing term, as shown in the next sections.

5.3.3.1 Rv scheme: positivity and energy stability

We start by rewriting (5.33) as

$$\begin{aligned} \phi_i^{\text{Rv}} &= \frac{1}{3} \sum_{j \in E} k_j u_j + \frac{1}{3} \alpha \sum_{\substack{j \in E \\ j \neq i}} (u_i - u_j) = \\ &= -\frac{1}{3} \sum_{\substack{j \in E \\ j \neq i}} k_j (u_i - u_j) + \frac{1}{3} \alpha \sum_{\substack{j \in E \\ j \neq i}} (u_i - u_j) = \\ &= \frac{1}{3} \sum_{\substack{j \in E \\ j \neq i}} (\alpha - k_j) (u_i - u_j) \end{aligned}$$

where (3.17) has been used in the second equality. Last expression shows that Rv scheme can be recast as in (4.1) with $3c_{ij}^E = (\alpha - k_j) \geq 0$ by definition of α . As a consequence, the scheme respects the sub-element LED condition, hence it verifies propositions 4.1.3, 4.1.4, 4.1.7, 4.1.8 and 4.1.5, and theorems 4.1.6 and 4.1.9, and the related stability bounds. In particular, the time-step restriction for its positivity reads

$$\Delta t \leq \frac{3|S_i|}{(1-\theta) \sum_{E \in \mathcal{D}_i} \sum_{\substack{j \in E \\ j \neq i}} (\alpha - k_j)} \quad \theta \in [0, 1). \quad (5.34)$$

Similarly, its local positivity is constrained by

$$\Delta t \leq \frac{|E|}{(1-\theta) \sum_{\substack{j \in E \\ j \neq i}} (\alpha - k_j)} \quad \theta \in [0, 1), \quad \forall i \in E \quad \forall E \in \mathcal{T}_h. \quad (5.35)$$

The Rv scheme is unconditionally positive when backward Euler time-integration is used in (5.5). As for the upwind $\mathcal{FV} - \mathcal{RD}$ scheme, the distribution coefficients of the Rv scheme are not guaranteed to be bounded, hence the scheme is not \mathcal{LP} .

The energy stability of the Rv scheme can be easily shown noting that

$$\frac{d\mathcal{E}^{\text{Rv}}}{dt} = - \sum_{i \in E} \sum_{j \in E} \frac{1}{3} u_i k_j u_j - \frac{1}{3} \sum_{i \in E} \sum_{j \in E} u_i \alpha (u_i - u_j) = \frac{d\mathcal{E}^{\text{C}}}{dt} - \epsilon^{\text{Rv}} \quad (5.36)$$

with the dissipation term reading

$$\epsilon^{\text{Rv}} = \frac{1}{3} \begin{bmatrix} u_1 - u_2 \\ u_1 - u_3 \\ u_2 - u_3 \end{bmatrix}^T \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix} \begin{bmatrix} u_1 - u_2 \\ u_1 - u_3 \\ u_2 - u_3 \end{bmatrix} \geq 0 \quad \text{if } \alpha \geq 0 \quad (5.37)$$

Proceeding as for the PG scheme, we obtain the energy estimate

$$\frac{d\mathcal{E}_h^{\text{Rv}}}{dt} = - \int_{\Omega} u_h \vec{a} \cdot \nabla u_h \, dx \, dy - \epsilon_h^{\text{Rv}} = - \frac{1}{2} \oint_{\partial\Omega} u_h (a \cdot \hat{n}) u_h \, dl - \epsilon_h^{\text{Rv}}$$

with

$$\epsilon_h^{\text{Rv}} = \sum_{E \in \mathcal{T}_h} \epsilon^{\text{Rv}}$$

Writing the boundary integral as

$$- \frac{1}{2} \oint_{\partial\Omega} u_h (a \cdot \hat{n}) u_h \, dl = - \frac{1}{2} \oint_{\partial\Omega} u_h |a \cdot \hat{n}| u_h \, dl - \oint_{\partial\Omega} u_h (a \cdot \hat{n})^- u_h \, dl$$

and using the BCs (5.31), we finally obtain the energy stability estimate

$$\frac{d\mathcal{E}_h^{\text{Rv}}}{dt} = - \frac{1}{2} \oint_{\partial\Omega} u_h |a \cdot \hat{n}| u_h \, dl - \epsilon_h^{\text{Rv}} \leq 0 \quad (5.38)$$

While the streamline dissipation of the PG scheme (5.25) has a residual character, since it is proportional to the residual through uniformly bounded coefficients, the dissipation of the Rv scheme is not residual, and it has a completely isotropic character.

5.4 Multidimensional Upwind schemes

We finally introduce the schemes which are at the basis of the developments of the thesis. They are built on ideas which are peculiar to the \mathcal{FS} framework and can hardly be incorporated in completely different set-ups. We consider here the so-called *multidimensional upwind* schemes defined as follows.

Definition 5.4.1. A \mathcal{FS} scheme is multidimensional upwind (\mathcal{MU}) if

- (i) in a 1-target element E , if $k_j > 0$ and $k_i, k_k < 0$, then: $\phi_j = \phi_h$ and $\phi_i = \phi_k = 0$
- (ii) in a 2-target element E , if $k_k < 0$ and $k_i, k_j > 0$, then: $\phi_k = 0$

Going back to the onedimensional analogy of figure 5.1, it is clear that \mathcal{MU} schemes reduce to 1D upwind schemes along the streamline cutting the triangle. In particular, all the information contained in the fluctuation is sent downstream to the outflow point. As a consequence, *all \mathcal{MU} schemes are equivalent in the one-target case*, in which the outflow point actually coincides with one of the points of the element. Conversely, in the two-target case, the information has to be split between the two nodes downstream. Clearly this is an important simplification from the point of view of the design of the schemes, since *different \mathcal{MU} schemes are defined just by choosing different distribution strategies in the two-target case*. Moreover, the geometrical framework described in figure 5.1 allows to perform this choice on the basis of heuristics making use of the directional propagation of the information which characterizes exact solutions. There is quite a number of possible choices one can make as shown in [129, 91, 148, 149]. Here we will present and analyze in more detail two of these possibilities, which probably have had the greatest success. However, before going into the details of the definition of these two schemes, we recall the following simple result.

Proposition 5.4.2 (\mathcal{MU} schemes, LED and \mathcal{LP} property: 1-target case). *In a 1-target element, \mathcal{MU} schemes are \mathcal{LP} and respect the sub-element LED condition.*

Proof. Let $(1, 2, 3)$ be the nodes of the 1-target triangle E and suppose 1 is the only downstream node: $k_1 > 0$, $k_2, k_3 < 0$. Linearity preservation is shown by

$$\beta_1 = 1, \beta_2 = \beta_3 = 0$$

which are clearly uniformly bounded. The local LED condition can be shown by noting that $\phi_2 = \phi_3 = 0$ while, due to (3.17) (see also (5.5) and (4.1))

$$\phi_1 = \phi^h = -k_2(u_1 - u_2) - k_3(u_1 - u_3) = c_{12}^E(u_1 - u_2) + c_{13}^E(u_1 - u_3)$$

with c_{12}^E and c_{13}^E positive by hypothesis. \square

Note that last proposition is not in contradiction with Godunov's theorem, since the LED property (hence positivity) and the sub-element LED property (hence local positivity) would require the positivity of the coefficients in all the elements of the mesh, which obviously are not all 1-target. Similarly, \mathcal{LP} schemes must have bounded coefficients in all $E \in \mathcal{T}_h$. Only one (or none) of the two properties (LED or \mathcal{LP}) can be retained in the two-target case.

5.4.1 The LDA scheme

The LDA (Low Diffusion A) is the most successful linear linearity preserving \mathcal{MU} scheme. It is defined by the following distribution coefficients:

$$\beta_i^{\text{LDA}} = k_i^+ N = k_i^+ \left(\sum_{j \in E} k_j^+ \right)^{-1} \geq 0. \quad (5.39)$$

In the homogeneous case, (5.11) gives for the local nodal residuals

$$\phi_i^{\text{LDA}} = \beta_i^{\text{LDA}} \phi^h = k_i^+ (u_{out} - u_{in}) \quad (5.40)$$

It is clearly \mathcal{LP} , since β_i^{LDA} remains bounded independently on ϕ^h , which implies $\phi_i = \mathcal{O}(h^3)$. One can also easily check that it does not respect the LED condition [129]. In the 2-target case, the LDA scheme admits a simple geometrical interpretation [129, 91, 148]. With reference to figure 5.4, we define the sub-triangles T_{423} and T_{143} . Simple trigonometry shows that

$$|T_{423}| = \frac{l_{34} k_1}{\|\vec{a}\|}, \quad |T_{143}| = \frac{l_{34} k_2}{\|\vec{a}\|} \quad \text{and} \quad |E| = |T_{423}| + |T_{143}| = \frac{l_{34}(k_1 + k_2)}{\|\vec{a}\|}$$

As a consequence, the distribution coefficients of the two downstream nodes 1 and 2 can be written as the area ratios

$$\beta_1^{\text{LDA}} = \frac{k_1}{k_1 + k_2} = \frac{|T_{423}|}{|E|}, \quad \beta_2^{\text{LDA}} = \frac{k_2}{k_1 + k_2} = \frac{|T_{143}|}{|E|}$$

The closer the outflow point is to node 1, the closer $|T_{423}|$ is to $|E|$ and $|T_{143}|$ to zero. Hence, the 1-target distribution is reached continuously with respect to the orientation of \vec{a} . This geometrical representation extends to the inhomogeneous case, in which the definition of the scheme remains unchanged but the residual is given by (5.14).

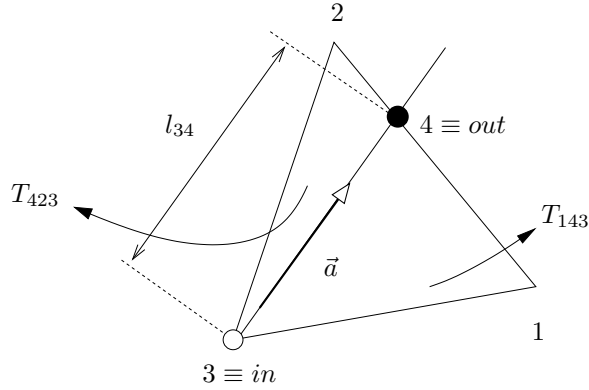


Figure 5.4: Geometry of \mathcal{FS} schemes. LDA in the 2-target case

5.4.1.1 LDA scheme: energy stability

Consistently with Godunov's theorem, the scheme does not respect any of the positivity (or LED) criteria discussed in chapter 4. The energy analysis of the scheme does not lead to a clear proof of stability [9, 4], however, it reveals an interesting form of the energy balance of the LDA scheme in the homogeneous case. In particular, if $(1, 2, 3)$ are the nodes of element E , using the definition of ϕ_i^{LDA} and expressions (5.7) and (5.11), we can write the local energy balance (4.36) as follows

$$\frac{d\mathcal{E}^{\text{LDA}}}{dt} = - (U^T (\mathcal{C}_{\text{LDA}} U) + (\mathcal{C}_{\text{LDA}} U)^T U) = -u_{\text{out}} \left(\sum_{j \in E} k_j^+ \right) (u_{\text{out}} - u_{\text{in}})$$

with $U = [u_1, u_2, u_3]^T$ and $(\mathcal{C}_{\text{LDA}})_{ij} = \beta_i^{\text{LDA}} k_j$. Simple manipulations lead to the more convenient expression

$$\begin{aligned} \frac{d\mathcal{E}^{\text{LDA}}}{dt} &= -\frac{1}{2} \overbrace{\left((u_{\text{out}} + u_{\text{in}}) \left(\sum_{j \in E} k_j^+ \right) (u_{\text{out}} - u_{\text{in}}) \right)}^{\text{centered scheme}} \\ &\quad - \frac{1}{2} \overbrace{\left((u_{\text{out}} - u_{\text{in}}) \left(\sum_{j \in E} k_j^+ \right) (u_{\text{out}} - u_{\text{in}}) \right)}^{\geq 0} \\ &= -\frac{(u_{\text{out}} + u_{\text{in}})}{2} \left(\sum_{j \in E} k_j^+ \right) (u_{\text{out}} - u_{\text{in}}) - \epsilon^{\text{LDA}} \end{aligned} \quad (5.41)$$

As for the PG and the Rv scheme, the energy balance reveals that the energy production of the LDA scheme can be split into a stabilizing term related to the dissipative mechanism of the multidimensional upwinding plus a centered term. However, in this case the central discretization acts along the streamline which renders the analysis less clear. Denoting by ζ_E the *segment* of streamline joining u_{in} and u_{out} and by ζ the stream-aligned coordinate running from u_{in} to u_{out} , we have that

$$\frac{(u_{\text{out}} + u_{\text{in}})}{2} \left(\sum_{j \in E} k_j^+ \right) (u_{\text{out}} - u_{\text{in}}) = \int_{\zeta_E} u_h a^* \frac{\partial u_h}{\partial \zeta} d\zeta, \quad a^* = \sum_{j \in E} k_j^+$$

so that, assembling the contributions from all the elements of the mesh, we get

$$\frac{d\mathcal{E}_h^{\text{LDA}}}{dt} = - \sum_{E \in \mathcal{T}_h} \int_{\zeta_E} u_h a^* \frac{\partial u_h}{\partial \zeta} d\zeta - \epsilon_h^{\text{LDA}} \quad (5.42)$$

with

$$\epsilon_h^{\text{LDA}} = \sum_{E \in \mathcal{T}_h} \epsilon^{\text{LDA}} \geq 0$$

The energy balance (5.42) shows clearly the analogy with the same expression for the PG scheme (5.30). However, in this case it is not clear at all how the first term could be reduced to a boundary integral, so that the BCs can be introduced into the analysis, eventually leading to an energy stability proof. Lastly, we remark that, as for the PG scheme, the dissipation term of the LDA scheme also has a residual character, due to the fact that the LDA scheme is \mathcal{LP} .

5.4.2 The N scheme

The N (Narrow) scheme is definitely the most successful first-order scheme designed for the solution of the advection equation. Firstly proposed by Roe [148], it has been since then the basis for the construction of \mathcal{LP} nonlinear positive discretizations. Moreover, thanks to its \mathcal{MU} character it has the lowest numerical dissipation among first-order schemes (see *e.g.* [126, 176]). It is defined by the following local nodal residuals:

$$\phi_i^N = k_i^+(u_i - u_{in}) . \quad (5.43)$$

Being \mathcal{MU} , the N scheme differs from the LDA scheme only in the 2-target case, in which it admits a simple geometrical representation. With reference to figure 5.5, we introduce the vectors \vec{a}_1 and \vec{a}_2 , parallel to the edges $\overline{31}$ and $\overline{32}$ respectively, such that $\vec{a}_1 + \vec{a}_2 = \vec{a}$. Simple algebra shows that

$$\phi^h(\vec{a}) = \int_E \vec{a} \cdot \nabla u_h \, dx \, dy = \phi^h(\vec{a}_1) + \phi^h(\vec{a}_2) = k_1(u_1 - u_3) + k_2(u_2 - u_3)$$

which immediately gives for the N scheme

$$\phi_1^N = k_1(u_1 - u_3) = \phi^h(\vec{a}_1), \quad \phi_2^N = k_2(u_2 - u_3) = \phi^h(\vec{a}_2)$$

In the 2-target case, the scheme reduces to first-order upwinding along the edges of the element meeting in the inflow point. As the $\mathcal{FV}-\mathcal{RD}$ and Rv schemes, the distribution coefficients of the N scheme can be unbounded as $\phi^h \rightarrow 0$, hence the scheme is not \mathcal{LP} .

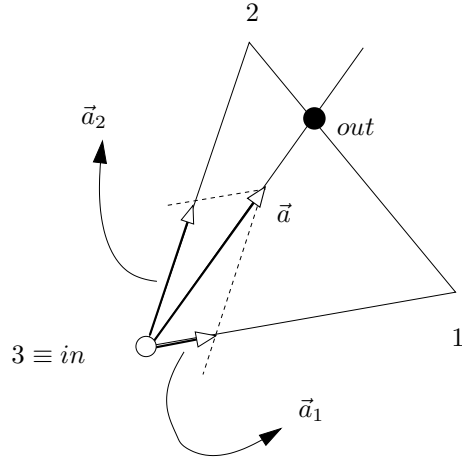


Figure 5.5: Geometry of \mathcal{FS} schemes. N in the 2-target case

5.4.2.1 N scheme: positivity and energy stability

The N scheme can be easily checked to respect the local LED condition:

$$\begin{aligned} \phi_i^N &= k_i^+ u_i + \sum_{j \in E} k_i^+ N k_j^- u_j = - \sum_{j \in E} k_i^+ N k_j^- u_i + \sum_{j \in E} k_i^+ N k_j^- u_j = \\ &= - \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N k_j^- (u_i - u_j) = \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i - u_j) \end{aligned}$$

with $c_{ij}^E = -k_i^+ N k_j^- \geq 0$. Hence, the N scheme verifies propositions 4.1.3, 4.1.4, 4.1.7, 4.1.8 and 4.1.5, and theorems 4.1.6 and 4.1.9, and the related stability bounds. In particular, the time-step restriction for its positivity reads

$$\Delta t \leq \frac{|S_i|}{(1-\theta) \sum_{E \in \mathcal{D}_i} k_i^+} \quad \theta \in [0, 1). \quad (5.44)$$

Similarly, its local positivity is constrained by

$$\Delta t \leq \frac{|E|}{3(1-\theta)k_i^+} \quad \theta \in [0, 1), \quad \forall i \in E | k_i > 0, \quad \forall E \in \mathcal{T}_h. \quad (5.45)$$

These constraints can be shown to be larger than the corresponding ones of the upwind $\mathcal{FV} - \mathcal{RD}$ scheme and of the Rv scheme [126, 129] (equations (5.21)-(5.22) and (5.34)-(5.35) respectively). The N scheme is unconditionally positive when backward Euler time-integration is used in (5.5). In addition to this, we note that

$$\sum_{\substack{j \in E \\ j \neq i}} (c_{ij}^E - c_{ji}^E) = - \sum_{\substack{j \in E \\ j \neq i}} (k_i^+ N k_j^- - k_j^+ N k_i^-) = - \sum_{j \in E} (k_i^+ N k_j^- - k_j^+ N k_i^-) = k_i^+ + k_i^- = k_i$$

which, as in the case of the upwind $\mathcal{FV} - \mathcal{RD}$ scheme cancels identically when summed over the elements of \mathcal{D}_i , in the case of constant advection. As a consequence the scheme respects the energy stability criteria of propositions 4.2.2 and 4.2.7 and proposition 4.2.8 in the fully discrete case. In particular, it can be easily shown that the equivalent local matrix energy operator of the N scheme is given by (see proposition 4.2.7 and [4, 18, 9])

$$\begin{aligned} \overline{M}^N &= \frac{1}{2} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} N \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix}^T + \\ &\quad \frac{1}{2} \begin{bmatrix} k_1^+ & 0 & 0 \\ 0 & k_2^+ & 0 \\ 0 & 0 & k_3^+ \end{bmatrix} - \frac{1}{2} \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix} N \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix}^T + \\ &\quad \frac{1}{2} \begin{bmatrix} -k_1^- & 0 & 0 \\ 0 & -k_2^- & 0 \\ 0 & 0 & -k_3^- \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -k_1^- \\ -k_2^- \\ -k_3^- \end{bmatrix} N \begin{bmatrix} -k_1^- \\ -k_2^- \\ -k_3^- \end{bmatrix}^T \quad (5.46) \end{aligned}$$

Boundary terms can be included in the analysis as in the case of the $\mathcal{FV} - \mathcal{RD}$ scheme and treated as done for the Galerkin and SUPG schemes.

5.4.3 Relations between the N and LDA schemes: dissipation

Before discussing nonlinear \mathcal{FS} discretizations, we will elaborate on the relations between the N and the LDA schemes in the 2-target case. The results presented here will be very useful later in this chapter and also in the following ones. Moreover, they lead us to a formulation of the N scheme which will be used in the inhomogeneous case and will be very important in the case of nonlinear \mathcal{CL} s. In particular, we want to show that in the 2-target case, the N scheme can be written as the LDA scheme plus an anisotropic dissipation term. To do this, we make the following observation. The definition of the inflow state (5.10) is such that, for the N scheme, one has automatically

$$\phi^h = \sum_{j \in E} \phi_j^N$$

However, as an exercise we can try to reverse things and, given the residual ϕ^h , compute u_{in} by requiring the satisfaction of the \mathcal{RD} consistency constraint (5.2). In formulas:

$$\sum_{j \in E} k_j^+ (u_j - u_{in}) = \phi^h \implies u_{in} = N \left(\sum_{j \in E} k_j^+ u_j - \phi^h \right). \quad (5.47)$$

Clearly, if ϕ^h is given by (5.7), using the relation $k_j = k_j^+ + k_j^-$, we get back (5.10). However, we can obtain additional information by using (5.47) in (5.43):

$$\phi_i^N = k_i^+ (u_i - u_{in}) = k_i^+ u_i - k_i^+ \overbrace{\sum_{j \in E} N k_j^+ u_j}^{=u_{out}} + k_i^+ \overbrace{N \phi^h}^{=\phi_i^{LDA}}$$

and finally

$$\phi_i^N = \phi_i^{LDA} + d_i^N, \quad d_i^N = k_i^+ (u_i - u_{out}) \quad (5.48)$$

Clearly, the term d_i^N is such that the local LED condition is verified, as shown in the previous section. Moreover, the definition of u_{out} ensures that

$$\sum_{j \in E} d_j^N = 0 \quad (5.49)$$

We can say more about this term if we write the local energy balance of the N scheme. Denoting by $(1, 2, 3)$ the nodes of the element, we define the vector $d^N = [d_1^N, d_2^N, d_3^N]^T$ given by:

$$d^N = D^N U, \quad D^N = \begin{bmatrix} k_1^+ & 0 & 0 \\ 0 & k_2^+ & 0 \\ 0 & 0 & k_3^+ \end{bmatrix} - \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix} N \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix}^T \quad (5.50)$$

The matrix D^N is clearly symmetric. Moreover, it is also positive semi-definite, as shown by the fact that

$$\begin{aligned} \epsilon^N = U^T D^N U = & (u_1 - u_2) k_1^+ N k_2^+ (u_1 - u_2) + \\ & (u_1 - u_3) k_1^+ N k_3^+ (u_1 - u_3) + \\ & (u_2 - u_3) k_2^+ N k_3^+ (u_2 - u_3) \geq 0 \end{aligned} \quad (5.51)$$

Last equation shows that the additional term d_i^N is indeed a dissipation term, in particular that *the N scheme is more dissipative than the LDA scheme* [3, 9]. In formulas (see equations (4.36), (5.41) and (5.48)):

$$\frac{d\mathcal{E}^N}{dt} = \frac{d\mathcal{E}^{\text{LDA}}}{dt} - \epsilon^N \leq \frac{d\mathcal{E}^{\text{LDA}}}{dt} . \quad (5.52)$$

Note that the dissipation terms d_i^N *do not* have a residual character and are quite anisotropic, acting essentially in the *cross-wind* direction.

5.4.4 An N scheme for inhomogeneous advection

The analysis of the last section gives a possible extension of the N scheme to the case $\mathcal{S}(x, y) \neq 0$. In particular, just by requiring the N scheme to be formally given by (5.43) also in the inhomogeneous case, and requiring that the \mathcal{RD} consistency condition (5.2) is verified with respect to the complete element residual (5.14), we obtain again equation (5.48). However, now the term ϕ_i^{LDA} also contains the terms coming from the integral of \mathcal{S} . Isolating these terms, we obtain

$$\phi_i^N = k_i^+(u_i - u_{in}) - \beta_i^{\text{LDA}} \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (5.53)$$

Finally, our construction has led us back to the N scheme proposed in [151, 160] for non-homogeneous advection, given precisely by (5.53). It respects by construction the consistency conditions (4.42) with

$$\beta_i^{\mathcal{S}} = \beta_i^{\text{LDA}}$$

Moreover, it can be recast as in (4.40) with

$$c_{ij}^{\mathcal{S}} = \beta_i^{\text{LDA}} \frac{|E|}{3} \geq 0 .$$

The positivity of these coefficients implies that the scheme verifies the hypotheses of propositions 4.3.1 and 4.3.2, and of theorem 4.3.3 and the related stability bounds. Note that the geometrical interpretation given for the LDA scheme also applies to this N scheme, as far as the distribution of the integral of the source term is concerned.

5.5 Nonlinear schemes

Nonlinear schemes are needed to combine linearity preservation and LED, as stated by theorem 4.4.5. The interest in \mathcal{FS} discretizations is largely due to the incredible success of the nonlinear PSI scheme of Struijs [165]. For steady scalar advection, in fact, the PSI scheme has been proved to be incredibly more accurate than second-order \mathcal{FV} schemes on irregular grids [165, 126, 129, 151, 160, 10, 9]. Its appeal is even greater considering

the fact that it is completely parameter free, thus better than \mathcal{FE} schemes with shock-capturing terms [126, 129, 39]. The problem is that, when dealing with inhomogeneous or time-dependent problems and, more importantly, systems of conservation laws, the extension of the PSI scheme is unclear. This has led to a large number of techniques to design nonlinear \mathcal{FS} schemes for which we refer to the discussion and to the references presented in chapter 1. Here we discuss two of these approaches: the local blending of a linear \mathcal{LP} scheme with a linear LED one, and the nonlinear *limiting* of a LED scheme into a \mathcal{LP} one. We mainly consider discretizations which use as linear LED scheme the N scheme, even though some comments are given on the use of Rv scheme for this purpose.

5.5.1 Blended schemes

The local blending of a linear \mathcal{LP} scheme with a linear LED scheme is one of the techniques more diffused in literature. The blending has to be nonlinear in order to be able to combine LED and linearity preservation. Given a \mathcal{LP} scheme defined by the split residuals $\phi_i^{\mathcal{LP}}$, and a linear first-order scheme verifying the sub-element LED condition, defined by the local nodal residuals ϕ_i^{LED} , a blended scheme is defined by

$$\phi_i = (1 - \Theta(u_h))\phi_i^{\mathcal{LP}} + \Theta(u_h)\phi_i^{\text{LED}} \quad (5.54)$$

where $\Theta(u_h)$ is a blending parameter, depending on the local structure of the numerical solution, which must ensure that $\phi_i = \mathcal{O}(h^3)$ in regions where u_h is *smooth*, at the same time ensuring in some way that the LED character of the first-order scheme prevails across discontinuities. Even though the idea is quite simple, the design of Θ is not trivial at all. As anticipated, here we mainly consider the case $\phi_i^{\mathcal{LP}} = \phi_i^{\text{LDA}}$ and $\phi_i^{\text{LED}} = \phi_i^{\text{N}}$, which is the one mostly used in published works on \mathcal{RD} (see [3, 155, 154, 55, 57, 86, 87] and references therein). In this case, the blending approach has an interesting interpretation. In particular, using (5.48) we can write that

$$\phi_i = (1 - \Theta(u_h))\phi_i^{\text{LDA}} + \Theta(u_h)\phi_i^{\text{N}} = (1 - \Theta(u_h))\phi_i^{\text{LDA}} + \Theta(u_h)\phi_i^{\text{LDA}} + \Theta(u_h)d_i^{\text{N}}$$

leading finally to

$$\phi_i = \phi_i^{\text{LDA}} + \Theta(u_h)d_i^{\text{N}} \quad (5.55)$$

Hence, blending the LDA and the N scheme is equivalent to adding to the LDA scheme a *nonlinear* dissipation term, proportional to the anisotropic *extra* dissipation of the N scheme (5.48). This is of course analog to what is done in \mathcal{FE} and, as in this case, presents the degree of freedom of the definition of the nonlinear dissipation, in the \mathcal{RD} case of $\Theta(u_h)$. Often, defining Θ in a very rigorous way, such that LED and linearity preservation are analytically proved, might not be extremely important in practice. This is shown by the fact that the *heuristic* definition of the blending parameter of Deconinck and collaborators [55, 154, 155, 57]

$$\Theta(u_h) = \frac{|\phi^h|}{\sum_{j \in E} |\phi_j^{\text{N}}|} \in [0, 1] \quad (5.56)$$

has given very good results in several fields of application [154, 87, 86, 9, 48, 52]. Note that as defined in (5.56), Θ only guarantees that $\phi_i = \mathcal{O}(h^3)$ in smooth regions, but there is no guarantee that the LED condition is verified, even for linear scalar advection. A rigorous study of this problem is found in [3, 9]. In particular, in the reference it is also shown that the PSI scheme of Struijs can be rewritten as a blended LDA/N scheme, for a particular choice of $\Theta(u_h)$.

5.5.1.1 Blended schemes and energy stability

When blending the LDA and the N scheme, the energy stability analysis benefits from the analysis of the LDA scheme. With the notation of equation (5.42), the energy balance is easily shown to be

$$\frac{d\mathcal{E}_h}{dt} = - \sum_{E \in \mathcal{T}_h} \int_{\zeta_E} u_h a^* \frac{\partial u_h}{\partial \zeta} d\zeta - \epsilon_h^{\text{LDA}} - \sum_{E \in \mathcal{T}_h} \Theta \epsilon^{\text{N}} \quad (5.57)$$

with ϵ^{N} given by (5.51). As for the LDA scheme, it is not clear how to treat the first term in the balance. A stable scheme could be obtained by blending the Rv scheme (5.33) with the PG scheme (5.26), giving the local nodal residual:

$$\phi_i = (1 - \Theta)\phi_i^{\text{PG}} + \Theta\phi_i^{\text{Rv}} = \phi_i^{\text{C}} + (1 - \Theta)\tau \frac{k_i}{2|E|} + \Theta d_i^{\text{Rv}}$$

Using (5.32) and (5.38), the energy balance for this scheme reads

$$\frac{d\mathcal{E}_h}{dt} = -\frac{1}{2} \oint_{\partial\Omega} u_h |\vec{a} \cdot \hat{n}| u_h dl - \sum_{E \in \mathcal{T}_h} (1 - \Theta) \epsilon^{\text{PG}} - \sum_{E \in \mathcal{T}_h} \Theta \epsilon^{\text{Rv}} \leq 0$$

as long as $\Theta \in [0, 1]$. Note however, that Θ normally is close to 1 only in correspondence of discontinuities, so that this blended scheme fits really into the \mathcal{FE} framework:

$$\phi_i = \text{central} + \text{energy stabilization} + \text{shock capturing}$$

A construction more *faithful* to the \mathcal{FS} philosophy will be described in the next section.

5.5.2 The PSI scheme: limited nonlinear schemes

As already said, the nonlinear PSI scheme of Struijs is the most successful \mathcal{RD} scheme ever designed. Its success is a consequence of the LED character of the scheme, together with its linearity preservation, compactness and with the fact that it is completely parameter free. Several generalizations of the scheme exist for scalar advection (see *e.g.* [129]). However, the most general formulation is obtained introducing the framework of the so-called *limited* schemes [129, 126, 9, 10, 12]. Consider a first-order linear \mathcal{FS} scheme, with split residuals ϕ_i^{LED} , verifying the sub-element LED condition. Suppose then to have a *continuous nonlinear mapping* $\varphi(x_0, x_1, x_2, x_3) : \mathbb{R}^4 \mapsto \mathbb{R}^3$ such that

$$\varphi(x_0, x_1, x_2, x_3) = x_0(y_1, y_2, y_3) \quad (5.58)$$

with

$$x_j = 0 \implies y_j = 0 \quad \forall j = 1, 2, 3 \quad (5.59)$$

$$x_j(x_0 y_j) \geq 0 \quad \forall j = 1, 2, 3 \quad (5.60)$$

$$|y_j| < \infty \quad \forall j = 1, 2, 3 \quad (5.61)$$

$$y_1 + y_2 + y_3 = 1 \quad (5.62)$$

A limited \mathcal{FS} scheme is obtained as

$$(\phi_1, \phi_2, \phi_3) = \varphi(\phi^h, \phi_1^{\text{LED}}, \phi_2^{\text{LED}}, \phi_3^{\text{LED}}) \quad (5.63)$$

The properties of such a scheme are determined by those of the mapping. In particular, (5.62) guarantees that the scheme verifies the consistency condition (5.2). Property (5.61), together with (5.58), and with the continuity of the mapping, guarantees that the scheme is \mathcal{LP} . Moreover, conditions (5.59) and (5.60) guarantee that, if $\phi^h \neq 0$, then if $\phi_j^{\text{LED}} = 0$ also $\phi_j = 0$, otherwise one has

$$\phi_j = x_0 y_j = \frac{x_0 y_j}{x_j} x_j = \alpha_j x_j = \alpha_j \phi_j^{\text{LED}} \quad \text{with} \quad \alpha_j = \frac{x_0 y_j}{x_j} \geq 0$$

hence, the resulting scheme also verifies the sub-element LED condition. There are quite a number of constructions leading to functions φ verifying (5.58)-(5.62). A review can be found in [9, 10, 12, 118]. In particular, starting from the N scheme, one obtains the PSI scheme of Struijs with the choice

$$\varphi(x_0, x_1, x_2, x_3) = \frac{x_0}{\sum_{j=1,3} (x_0 x_j)^+} ((x_0 x_1)^+, (x_0 x_2)^+, (x_0 x_3)^+) \quad (5.64)$$

This formulation of the PSI scheme is known since long and it has been used to construct limited \mathcal{LP} nonlinear variants of the upwind $\mathcal{FV} - \mathcal{RD}$ scheme (5.20) already in [129, 126]. However, only lately this more general framework has emerged as a way of constructing nonlinear schemes for time-dependent problems and systems [10]. In this thesis, we make use of nonlinear schemes obtained applying (5.64) to the N scheme, and the resulting scheme will be referred to as the limited N scheme. However, as we will see, depending on the application, the definition of ϕ_i^{N} and of ϕ^h will change. We also remark that (5.64) can be recast in the simpler form. In particular, for the distribution coefficients of the limited scheme one can write

$$\beta_i = \frac{\max(0, \beta_i^{\text{N}})}{\sum_{j=1,3} \max(0, \beta_j^{\text{N}})}, \quad \beta_j^{\text{N}} = \frac{\phi_j^{\text{N}}}{\phi^h} \quad (5.65)$$

which is how the limited N scheme is normally presented in literature [129, 126, 10, 12]. Compared to the blending approach, the nonlinear mapping has the advantage of requiring only the evaluation of the local nodal residuals of the linear LED scheme. Generally, mapping (5.64) works quite well in a large number of cases. The results of this thesis confirm this quality. On very hard computations often the limited schemes work quite well where the blended schemes fail [50, 9, 10]. However, it must be also remarked that this mapping is known since a very long time and that improved constructions, leading to schemes *significantly* different from the ones obtained with (5.64)

still have to appear. It is expected that the study and the understanding of these non-linear mappings might be one of the most important subjects of future research in the \mathcal{FS} community. Here we will give some conditions for the well-posedness of the procedure, which were already underlined in [142, 141]. Finally, we discuss the issue of the energy stability of these schemes, which, for the time being, seems to represent the weak point of the whole construction.

5.5.2.1 Well-posedness of the mapping

We recall here the conditions for the well-posedness of the limiting procedure, as presented in [142, 141]. The construction discussed in the previous section seems to be quite flexible, in the sense that, once a linear LED scheme is available and *fed* to the mapping, one obtains a LED and \mathcal{LP} scheme. The only important property of the linear scheme is, from this point of view, the sign of its split residuals. Actually this is not quite accurate. As we are going to show, the linear LED scheme has to satisfy more conditions for the mapping to work. After discussing these conditions, we will also give examples of schemes proposed in literature which try to violate them.

Consider then a linear first-order scheme respecting the sub-element LED condition. Denote by ϕ_i^{LED} its local nodal residuals. Starting from this scheme, we want to find a nonlinear mapping satisfying properties (5.58)-(5.62) and use it to construct a nonlinear \mathcal{LP} scheme which also respects the sub-element LED condition. Due to (5.58), we can study the case $\phi^h \neq 0$ since if $\phi^h = 0$ we know that all the ϕ_i s are zero for the nonlinear scheme. Next we assume that

$$\sum_{j \in E} \phi_j^{\text{LED}} = \phi^1 \quad (5.66)$$

with ϕ^1 not necessarily equal to ϕ^h . Even though for the scalar advection equation considered in this chapter this seems unlikely to happen, in general this is possible. Even in the simple case of pure advection, an example of such a situation is given by the \mathcal{FV} scheme (5.18) in the case $\vec{a} = \vec{a}(x, y)$. In fact, if $\nabla \cdot \vec{a} = 0$, the finite volume scheme gives

$$\phi^1 = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n} \, dl = \oint_{\partial E} (\vec{a}(x, y)u)_h \cdot \hat{n} \, dl$$

with a continuous piecewise linear approximation of the flux \mathcal{F}_h , which does not necessarily lead to (5.7) obtained by using the exact average of $\vec{a}(x, y)$ (equation (5.8)).

To derive conditions for the well-posedness of the limiting we recall here the properties that the mapping has to satisfy. Since $\phi^h \neq 0$ we can introduce the quantities

$$\beta_j^{\text{LED}} = \frac{\phi_j^{\text{LED}}}{\phi^h}$$

We then rewrite (5.58)-(5.62) as follows:

$$\varphi(\phi^h, \phi_1^{\text{LED}}, \phi_2^{\text{LED}}, \phi_3^{\text{LED}}) = \phi^h(\beta_1, \beta_2, \beta_3) \quad (lp0)$$

$$|\beta_j| < \infty \quad \forall j = 1, 2, 3 \quad (lp1)$$

$$\phi_j^{\text{LED}} = 0 \Rightarrow \beta_j = 0 \quad \forall j = 1, 2, 3 \quad (p0)$$

$$\beta_j^{\text{LED}} \beta_j \geq 0 \quad \forall j = 1, 2, 3 \quad (p1)$$

$$\beta_1 + \beta_2 + \beta_3 = 1 \quad (c)$$

where, conditions (lp0)-(lp1) guarantee that the nonlinear scheme is \mathcal{LP} , (c) is the consistency condition and (p0) and (p1) guaranteeing the local LED condition for the nonlinear scheme. Note that (p1) is obtained by dividing (5.60) by $(\phi^h)^2 > 0$. We now observe that, since condition (c) requires the sum of the nonlinear distribution coefficients to be equal to 1, we must have *at least* one $\beta_j > 0$. To have this, due to the positivity conditions (p0)-(p1), we must have at least one linear distributions coefficient $\beta_j^{\text{LED}} > 0$. However, due to (5.66), these coefficients satisfy

$$\sum_{j \in E} \beta_j^{\text{LED}} = \sum_{j \in E} \frac{\phi_j^{\text{LED}}}{\phi^h} = \frac{\sum_{j \in E} \phi_j^{\text{LED}}}{\phi^h} = \frac{\phi^1}{\phi^h} = \gamma$$

If $\gamma > 0$, then at least one of the β_j^{LED} s must be greater than zero. This ensures that the consistency condition (c) can be satisfied. Conversely, if $\gamma \leq 0$, we are likely to encounter the unfortunate situation in which

$$\beta_j^{\text{LED}} \leq 0 \quad \forall j \in E$$

In this case, we will not be able to satisfy (c), *unless the positivity conditions (p0)-(p1) are relaxed*. This proves that

Proposition 5.5.1 (Well-posedness of the mapping - sufficient condition).

Given a linear scheme satisfying the sub-element LED condition, a condition to construct a well-posed nonlinear mapping satisfying properties (5.58)-(5.62) is that

$$\phi^h \sum_{j \in E} \phi_j^{\text{LED}} > 0 \quad (5.67)$$

Clearly, we also have the (trivial) corollary

Corollary 5.5.2 (Well-posedness of the mapping). *Given a linear scheme satisfying the sub-element LED condition, a sufficient condition to construct a well-posed nonlinear mapping satisfying properties (5.58)-(5.62) is that*

$$\sum_{j \in E} \phi_j^{\text{LED}} = \phi^h \quad (5.68)$$

While one can ensure that the necessary condition holds, simply by requiring the linear LED scheme to respect the \mathcal{RD} consistency condition, the sufficient condition of proposition 5.5.2 is in general impossible to ensure and can only be checked during the simulation. Note also that we made no assumptions on the sign of the mapped distribution coefficients β_j . All the constructions presented in [10, 9, 118, 12] instead assume that $\beta_j \geq 0$, which makes our analysis even more important.

5.5.2.2 Well-posed mappings: a counterexample

We give an example showing the importance of the analysis. In particular, we consider the schemes proposed in [12] for the solution of the advection equation. In the reference, the authors introduce a framework that allows to extend the accuracy of \mathcal{FS} schemes to more than second-order. This is achieved by introducing piecewise quadratic and cubic approximations of the unknown, instead of the piecewise linear (3.7).

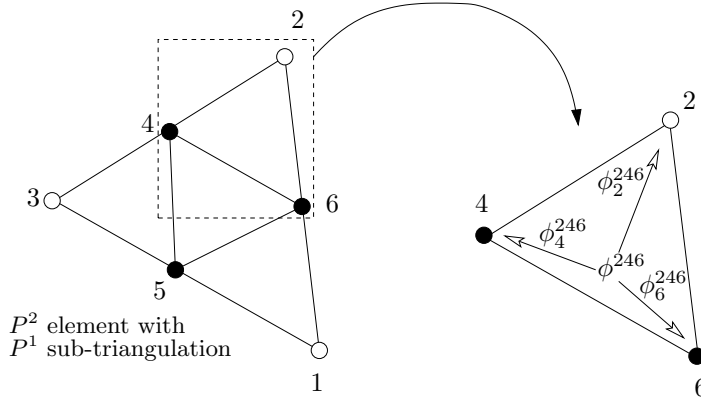


Figure 5.6: P2 element (left) and sub-element distribution (right)

Without going into too much detail, consider the case of a parabolic interpolation of the unknown, obtained by adding in each triangle of the mesh, the mid-points of the edges of the element as unknowns (see figure 5.6). In this way, one obtains a mesh of P^2 elements over which u_h can be expressed as in (3.7), except that now the basis functions ψ_i are parabolas. To obtain their schemes, in [12] the authors introduce a sub-triangulation of these 6-nodes elements as shown in figure 5.6. On each sub-triangle, the residual ϕ^h is computed according to (5.1) (with $\mathcal{S} = 0$), using the local parabolic representation of u_h to have the increased accuracy. As before, one can show that third-order schemes can be obtained by distributing ϕ^h with bounded coefficients. In order to also have a LED-type scheme, they use the limiting technique described in these pages. Let us consider, for example, what happens in the sub-triangle E^{246} depicted on the right on figure 5.6. The local residual of this sub-element is [12]

$$\int_{E^{246}} \vec{a} \cdot \nabla u_h \, dx \, dy = \sum_{j=1}^6 \int_{E^{246}} \vec{a} \cdot \nabla \psi_j u_j \, dx \, dy$$

Hence, due to the parabolic interpolation ϕ^{246} cannot be written using an expression similar to (5.7), involving only nodes 2, 4 and 6. As a consequence, one does not know how to design on E^{246} a scheme similar to the N scheme (5.43), respecting the sufficient condition (5.68). The authors have instead applied the limiting to the first-order N scheme constructed on E^{246} as if the unknown would vary linearly between the local values (u_2, u_4, u_6) . This case, then, fits precisely into the analysis made here. In particular, using (5.43) and (5.7), we have

$$\phi_2^{N,246} + \phi_4^{N,246} + \phi_6^{N,246} = k_2^{246}u_2 + k_4^{246}u_4 + k_6^{246}u_6 = \phi^1$$

where the k_j^{246} parameters are computed using the local geometry of E^{246} , and with $\phi^1 \neq \phi^{246}$ in general, since ϕ^1 corresponds to (5.1) computed with a local linear interpolation, while ϕ^{246} is obtained using the parabolic interpolation over the 6-nodes element. In particular, the authors of [12] apply to this *local* N scheme the mapping of the PSI given by (5.65). The analysis performed here shows that, Since the necessary condition (5.67) cannot be ensured *a priori* and since (5.67) does not hold, the nonlinear scheme obtained in this way cannot satisfy the consistency condition (c) (or equivalently (5.62)) and the LED condition (p1) (or equivalently (5.60)) at the same time, everywhere on the mesh. Indeed, the authors remark that the higher accuracy could not be observed unless (5.65) was modified as follows:

$$\beta_i = \frac{\max(0, \beta_i^N) + \epsilon}{\sum_j \max(0, \beta_j^N) + 3\epsilon}$$

with ϵ a small number set to 10^{-10} in their calculations. This agrees perfectly with our analysis. Going back to the sub-triangle E^{246} , we can observe that: if $\phi^1 \phi^{246} \leq 0$ with $\phi_j^{N,246} \phi^{246} \leq 0 \forall j = 2, 4, 6$, application of (5.65) leads to $\beta_j^{246} = 0 \forall j = 2, 4, 6$ for the nonlinear scheme, even if both ϕ^{246} and ϕ^1 are non-zero. This introduces an inconsistency in the scheme eventually spoiling its convergence with mesh refinement. Conversely, applying the modified formula, the nonlinear scheme reduces in this case to the central one which is consistent, however it is not LED. Hence, consistency has been recovered but only relaxing the LED conditions, as predicted by our analysis. This shows that, the extension of the limiting technique to more complex situations needs a better understanding of the properties of the mappings.

5.5.2.3 Limited schemes and energy stability

Energy stability seems to be a weak point of the limited schemes. Their inherent complexity makes a general analytical study very hard. To our knowledge, the only existing result on the topic is due to Barth [18], and it is not a positive result. In the reference, the PSI scheme of Struijs is analyzed in the simplest case: constant scalar advection. In this case, the analysis is quite simple and will be briefly recalled hereafter. Suppose to be in a 2-target element E. Let 1 and 2 be the downstream nodes. Hence, $k_1, k_2 > 0$ and $k_3 < 0$. Suppose also that $\phi^h > 0$. The PSI scheme is obtained by applying (5.65) to the N scheme (5.43). Obviously, since $\phi_i^N + \phi_2^N = \phi^h$, two things can happen. Either both ϕ_i^N and ϕ_2^N are positive, in which case one easily shown that

(5.65) gives back the N scheme, or one of the N scheme local residuals is negative. Suppose, for example, that $\phi_2^N < 0$. In this case, the limiting will send all the residual to node 1. We can then look at the local energy production of the PSI scheme (4.36). One easily shows that:

$$\frac{d\mathcal{E}^{\text{PSI}}}{dt} = -\frac{1}{2} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}^T \begin{bmatrix} 2k_1 & k_2 & k_3 \\ k_2 & 0 & 0 \\ k_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Using the energy equivalence lemma 4.2.3, we can study the energy of the equivalent operator:

$$\frac{d\mathcal{E}^{\text{PSI}}}{dt} = -\frac{1}{2} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}^T \begin{bmatrix} 2k_1 & k_2 & k_3 \\ k_2 & 0 & 0 \\ k_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}^T \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

which gives

$$\frac{d\mathcal{E}^{\text{PSI}}}{dt} = -\frac{1}{2} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}^T \begin{bmatrix} k_1 & k_2 & k_3 \\ k_2 & -k_2 & 0 \\ k_3 & 0 & -k_3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Recall that for constant scalar advection the terms added vanishes identically when assembling the contributions of all the elements. Straightforward calculations lead to

$$2\frac{d\mathcal{E}^{\text{PSI}}}{dt} = -k_1(u_1 - u_3)^2 - k_2((u_1 - u_3)^2 - (u_1 - u_2)^2)$$

which shows that the scheme is stable if one of the following conditions is verified

1. [Necessary condition] $k_1(u_1 - u_3)^2 + k_2((u_1 - u_3)^2 - (u_1 - u_2)^2) > 0$
2. [Sufficient condition] $|u_1 - u_3| > |u_1 - u_2|$

Note that the relations $\phi_1^N \phi^h > 0$, $\phi_2^N \phi^h < 0$, and $\phi^h > 0$, assumed by hypothesis, are not enough to guarantee neither the sufficient nor the necessary condition for the stability of the scheme [18]. Hence, when the limiting is reducing the number of nodes to which the residual is distributed, sources of energy instability might be introduced. As remarked in [18], this problem seems not to spoil the stability of the scheme in the scalar case, in which the scheme shows no problems of convergence toward the steady-state. A partial explanation of this can be that, for scalar problems, the PSI scheme is equivalent to the blended scheme of [3], which has been shown to respect the energy balance (5.57). Although (5.57) is not a real stability estimate, it shows that there is indeed a dissipation mechanism which is associated to the upwinding. This analysis is partially confirmed by the fact that the limited scheme obtained by applying (5.65) to the Rv scheme (5.33) converges very poorly to steady-state, as shown in [10, 9, 118]. Since Rusanov's scheme is not upwind, it is most of the times a 3-target scheme, in the sense that it distributes the residual to all three nodes of the element. In this case, it is more likely that the limiting will *compress* the distribution to two or even only one

node, probably introducing a destabilizing mechanism similar to the one seen for the PSI scheme. This feature seems to be particularly important for systems.

We remark that in this respect nonlinear limited \mathcal{RD} schemes are substantially different from stabilized Galerkin \mathcal{FE} schemes with nonlinear shock-capturing (\mathcal{SC}). Indeed, these \mathcal{SC} terms have by construction a dissipative character. Hence, the energy stability of the resulting schemes is quite clear, as the analysis of [19] shows. Nevertheless, the L^∞ -stable character of nonlinear \mathcal{FE} schemes in presence of discontinuities is only recovered globally, thanks to the regularization of the solution introduced by the additional nonlinear dissipation [166]. Conversely, nonlinear limited \mathcal{RD} schemes are constructed by imposing their local positivity. This guarantees the preservation of the local monotonicity of the solution. However, a dissipative character can only be achieved if the overall discretization maintains a marked upwind character. The \mathcal{RD} nonlinear limiting and the \mathcal{FE} nonlinear \mathcal{SC} are then two completely different approaches to stabilize discontinuities. The first has a strong L^∞ *flavor*, while the second relies on a very strong L^2 stabilization due to dissipation.

5.6 Illustrative examples

We show in this section some *sample* numerical results illustrating *experimentally* the properties of some of the scheme presented. In particular, we will consider the solution of the steady limit of the rotational advection equation

$$\frac{\partial u}{\partial t} + (y, 1 - x) \cdot \nabla u = \mathcal{S}(x, y) \quad \text{on } [0, 1] \times [0, 1] \subset \mathbb{R}^2$$

We present results on a homogeneous problem and on a non-homogeneous one. In both cases, we discretize the spatial domain $[0, 1] \times [0, 1]$ with an irregular triangular mesh (see figure 3.1) and use explicit FE Euler time-stepping (equation (4.6)) to march toward the steady-state. Local time-stepping has been used, computing the time-step as (see equations (4.9) and (5.44))

$$\Delta t_i = 0.9 \frac{|S_i|}{\sum_{E \in \mathcal{D}_i} k_i^+} \quad \forall i \in \mathcal{T}_h$$

We discuss the results obtained with the LDA scheme (5.39), with the N scheme (5.48) and with the limited variant of the N scheme obtained with mapping (5.65), which we refer to as the LN scheme.

5.6.1 Rotational advection: homogeneous case

We consider the homogeneous advection of the inlet profile given by

$$u(x, y = 0) = u_0(x) \begin{cases} \cos^2(\pi(x - 0.3)/0.4) & \text{if } 0.1 \leq x \leq 0.5 \\ 1 & \text{if } 0.7 \leq x \leq 0.9 \\ 0 & \text{otherwise} \end{cases}$$

The spatial domain is discretized with an irregular grid with reference element size $h = 1/50$. The exact solution can be computed with the method of characteristics (see §2.1) and can be easily seen to be given by

$$u = u(r, \theta) = u_0(r) \quad \forall \theta \quad \text{with} \quad \begin{cases} r = \sqrt{(x-1)^2 + y^2} \\ x = 1 - r \cos \theta \\ y = r \sin \theta \end{cases}$$

The computation is started setting as initial solution $u = 0$ everywhere, except on the lower boundary $y = 0$, where the exact solution is imposed. We plot on the left on figure 5.7 the convergence histories of the L^1 norm of the spatial residual for the N, LDA and LN scheme. All the schemes converge quickly to steady-state. On the right in the same figure we compare the profile of the numerical solutions at the outlet boundary $x = 1$ with the exact one: the LDA scheme reproduces perfectly the smooth part of the solution while giving oscillations in correspondence of the discontinuity; the N scheme reproduces poorly both parts of the solution, however it does not show any sign of oscillations; the limited N scheme gives both a good approximation of the smooth profile and a sharp and monotone discontinuity.

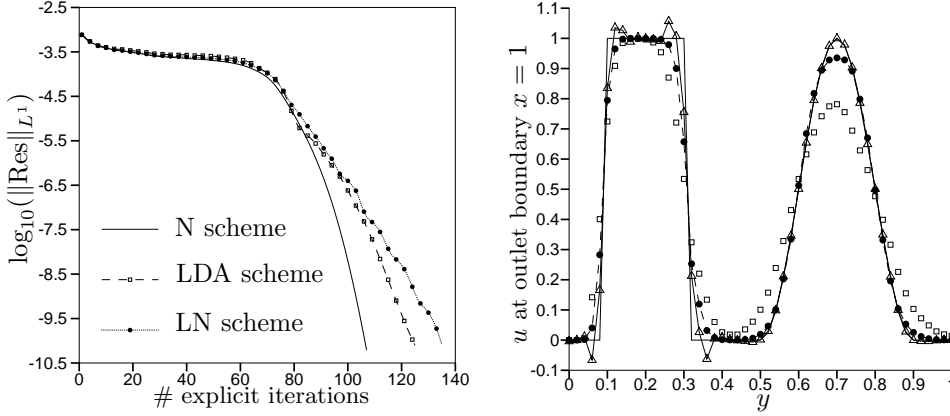


Figure 5.7: Homogeneous advection. Left: convergence histories. Right: solution at outlet boundary $x = 1$. Solid line: exact. Solid line with triangles: LDA scheme. Squares: N scheme. Dashed line with black circles: LN scheme

5.6.2 Rotational advection with a source term

We discuss here a more interesting case involving a discontinuous source term

$$\mathcal{S}(x, y) = \begin{cases} 10 & \text{if } r \leq 0.25 \\ 0 & \text{otherwise} \end{cases}, \quad r = \sqrt{(x-0.5)^2 + (y-0.3)^2}$$

On the inflow boundary $y = 0$ we set the solution to

$$u(x, y = 0) = u_0(x) \begin{cases} -5 & \text{if } 0.3 \leq x \leq 0.8 \\ 0 & \text{otherwise} \end{cases}$$

The exact solution, computed using the method of characteristics, has a somewhat complex structure. due to the interaction of the inlet profile with the source term. For this reason, an irregular grid with $h = 1/100$ has been used in the computations. A contour plot of the exact solution on the mesh is reported on the left in figure 5.8. The computations have been started by setting as initial solution $u = 0$ everywhere, except on the inflow boundary. The convergence histories of the LDA, N and LN scheme are reported on the right on figure 5.8. No convergence problems are encountered. The solution obtained with the LDA scheme is reported on figure 5.9. On the left picture we see the contour plot of the unknown u . The general structure of the solution is correct, however we clearly see the formation of numerical oscillations, as a consequence of the non-LED character of the scheme. On the right picture we compare the solution profile at $x = 1$ with the exact one. As before, while smooth regions of the solution are very well reproduced, in correspondence of sharp fronts spurious oscillations appear. We then report, on figure 5.10, the solution obtained with the N scheme. As in the homogeneous case, this scheme gives a quite poor reproduction of the structure of the exact solution. Sharp fronts are smeared over quite a few cells. However, the N scheme gives a monotone approximation of the unknown also in vicinity of discontinuities.

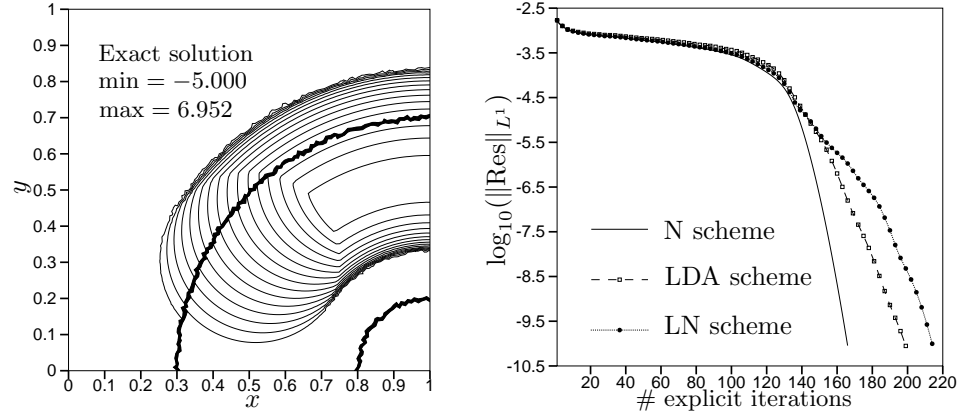


Figure 5.8: Non-homogeneous advection: exact solution (left) and convergence (right)

These observations are confirmed by looking at the right picture on figure 5.10, where the numerical solution at $x = 1$ is compared with the exact one: the smooth region is poorly resolved and jumps are smeared over several cells. Nevertheless, no oscillations whatsoever are present in the results, confirming the theoretical analysis.

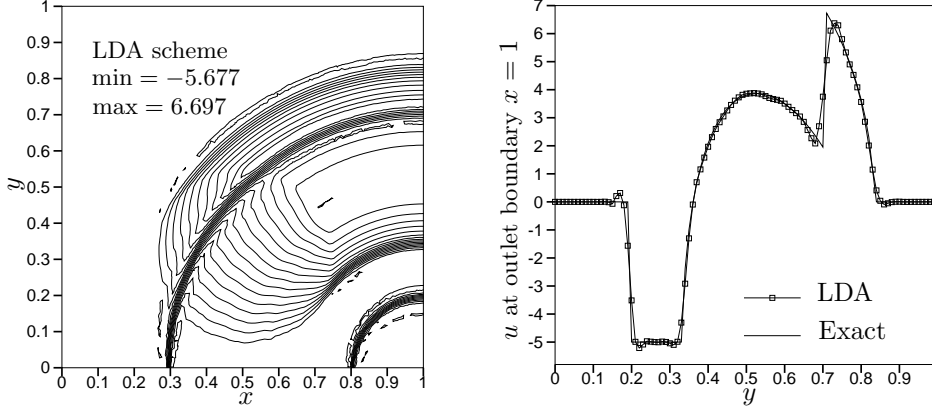


Figure 5.9: Non-homogeneous advection: LDA scheme. Left: contour plot of the solution. Right: solution at the outlet boundary $x = 1$

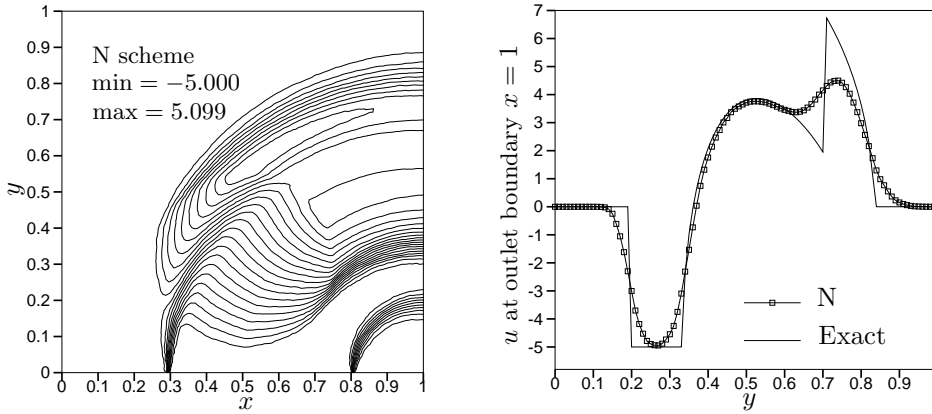


Figure 5.10: Non-homogeneous advection: N scheme. Left: contour plot of the solution. Right: solution at the outlet boundary $x = 1$

At last we report, in figure 5.11, the results obtained with the limited N scheme. From the contour plot on the left picture we can see that the scheme gives a smooth and accurate reproduction of the exact solution, without any sign of numerical oscillations. The distribution of u at the outflow boundary $x = 1$ confirm this. Discontinuities are kept quite sharp and absolutely no oscillations can be seen. The smooth regions are well reproduced, perhaps not as well as with the LDA scheme, but this is a price we can pay to gain the satisfaction of the discrete maximum principle. Overall, the solution is quite accurate and clean.

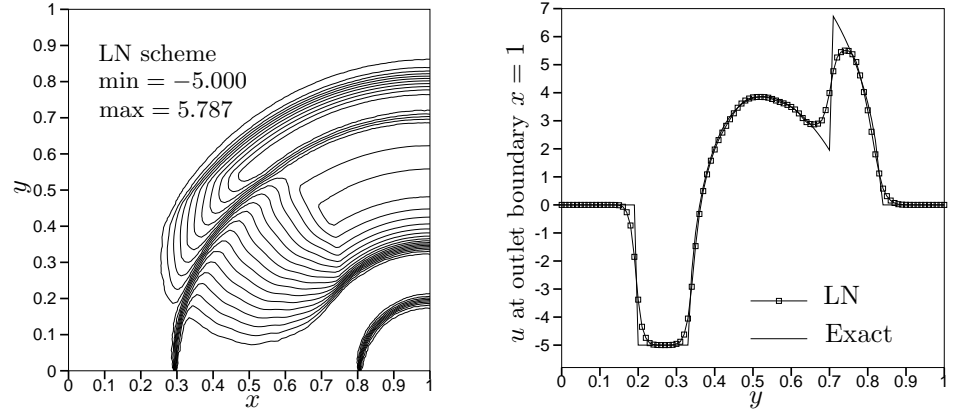


Figure 5.11: Non-homogeneous advection: LN scheme. Left: contour plot of the solution. Right: solution at the outlet boundary $x = 1$

5.7 Summary

This chapter has finally considered the definition of schemes for the solution of the advection equation. After defining in a general way what we intend for a \mathcal{FS} scheme, an overview has been given of several particular ways of deriving the discrete equations. The main results of the chapter can be summarized as follows.

- A less abstract solution procedure for the advection equation has been given. The schemes considered all fit into the abstract framework of (4.1). In particular, we have shown that not only \mathcal{FS} schemes can be modeled by this prototype but also first-order \mathcal{FV} schemes on the median dual cell and \mathcal{FE} schemes;
- The definition of a \mathcal{RD} scheme has allowed to define a particular type of second-order discretizations: the linearity preserving schemes. This concept will be always at the basis of the construction of high-order \mathcal{FS} schemes;
- An overview of schemes has been given showing how, in practice, the theoretical properties discussed in chapter 4 are checked for each of them;
- The concept of multidimensional upwinding has been introduced. Examples of \mathcal{MU} schemes have been given: the LDA and the N scheme;
- Manipulating the local energy balance of \mathcal{MU} schemes we have shown that the multidimensional upwinding indeed introduces a dissipation mechanism. However, even though we are able to write the energy balance in a way formally similar to what is done for PG finite element schemes with streamline dissipation, the *non-conventional* character of the \mathcal{MU} dissipation mechanism does not allow to derive a precise stability estimate;
- The relations between the LDA and N schemes, in terms of their dissipative character, have been underlined;

- The construction of nonlinear schemes has been analyzed. Blended schemes and limited schemes have been considered.
- The construction of limited nonlinear schemes has been studied, showing that for its well-posedness the underlying linear positive scheme must satisfy some *local consistency* with the element residual;
- The issue of the energy stability of nonlinear \mathcal{RD} schemes has been reviewed. The analysis of the stability of the limited schemes shows that sources of energy instability can be introduced by the mapping procedure. However, we have also shown that the \mathcal{MU} character of these schemes, and of the blended schemes as well, introduces a dissipative mechanism that, for scalar problems, seems to dominate. Hence, upwinding is one of the keys to stability;
- Illustrative computational examples have been given, confirming the theoretical analysis of this chapter and of chapter 4, both in the case of homogeneous and inhomogeneous advection.

Chapter 6

Nonlinear scalar conservation laws: \mathcal{CRD} schemes

This chapter is devoted to the extension of the \mathcal{FS} schemes presented in chapter 5 to the solution of the steady limit of scalar nonlinear conservation laws of the type:

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathcal{F}(u) = 0 \quad \text{on } \Omega \subset \mathbb{R}^2 \quad (6.1)$$

As seen in chapter 2, the solution of such a nonlinear problem has to be defined in a weak sense, due to the fact that, even if the initial and boundary data are smooth, nonlinear conservation laws evolve discontinuous solutions in a finite time. We will start with a motivational example to explain how this translates into a design criterion for the numerical schemes. This example will allow to introduce the concept of a *conservative discretization*. The first half of the chapter is then devoted to the presentation and the analysis of \mathcal{RD} schemes for nonlinear conservation laws which satisfy this criterion. The second half of the chapter instead considers the stability of the conservative \mathcal{FS} scheme presented, with respect to their dissipative character which is related, for a nonlinear conservation law, to the satisfaction of a discrete entropy inequality.

6.1 Conservative and non-conservative schemes

As a motivational example, consider the solution of (6.1) with the *exponential* flux

$$\mathcal{F}(u) = (e^u, u) .$$

In particular, consider the case in which $\Omega = [-0.025, 1.2] \times [0, 0.5]$ with BCs:

$$\begin{aligned} u(x, y=0) &= \begin{cases} \sin(2\pi x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \\ u(-0.025, y) &= 0 \end{aligned} \quad (6.2)$$

On a fine unstructured discretization of Ω ($h = 1/200$), we computed a reference steady solution of this problem with a nonlinear scheme we will describe in the following sections. A contour plot of this solution is reported in figure 6.1.

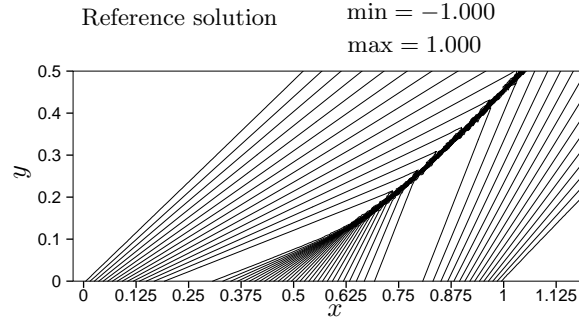


Figure 6.1: Nonlinear \mathcal{CL} with *exponential flux*. Contours of the reference solution

From the plot, we see that, even if the boundary data are continuous, with piecewise continuous derivatives, the solution contains a *shock* which develops at a finite and relatively small distance from the lower boundary, where the smooth data are imposed. Suppose now that we are going to solve this problem with one of the \mathcal{RD} schemes presented in chapter 5. For example we might want to use the limited N scheme, in order to achieve a monotone capturing of the shock. To do this, we have to be able to apply the N scheme (5.43) and then use mapping (5.65). Equation (6.1) is formally different from the advection equation, however, we can write it in a similar form, introducing the Jacobian of the flux $\vec{a}(u)$:

$$\frac{\partial u}{\partial t} + \vec{a}(u) \cdot \nabla u = 0, \quad \vec{a}(u) = \frac{\partial \mathcal{F}(u)}{\partial u} = (e^u, 1)$$

This equation could be treated as an advection equation with varying advection speed. As observed in §5.1.1, we can locally average this speed and then apply the schemes, as they have been described in chapter 5. In the generic element E of the mesh, suppose then to average the speed as

$$\bar{a} = \frac{1}{3} \sum_{j \in E} \vec{a}(u_j)$$

which is a second-order accurate approximation of (5.8). After having locally linearized the problem, we can apply the \mathcal{FS} schemes described in the previous chapter, computing the element residual as (5.7) and then distributing it using the limited N scheme. Since we are not using the conservative form of the problem in the discretization but

its quasi-linear form, we refer to this scheme as to the non-conservative LN scheme, or LN-NC scheme. The contour plot of the steady solution obtained in this way is reported on the left in figure 6.2. The result looks indeed very similar to the reference solution. However, a closer inspection will reveal that some important differences are present. In particular, on the right in figure 6.2, we have reported a close up of the shock in correspondence of the upper boundary.

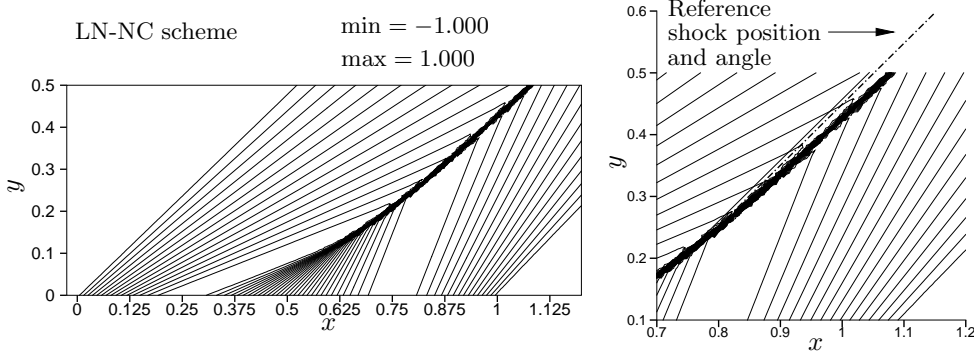


Figure 6.2: Nonlinear \mathcal{CL} with *exponential flux*. Contour plot of the solution obtained with the non-conservative LN scheme (left). Wrong shock angle (right)

Next to the result obtained with the non-conservative LN scheme, we have plotted a line indicating the position and the angle of the shock in the reference solution. Even though the same mesh has been used to compute these results, the LN-NC scheme seems to mispredict these features. This is confirmed if one looks at the profile of u along the boundary $y = 0.5$. In particular, on figure 6.3 we compare with the reference the distribution of the unknown along this line, computed with the LN-NC scheme on a coarser mesh ($h = 0.015$) and on the fine mesh ($h = 1/200$). We clearly see that the LN-NC scheme consistently gives a wrong result.

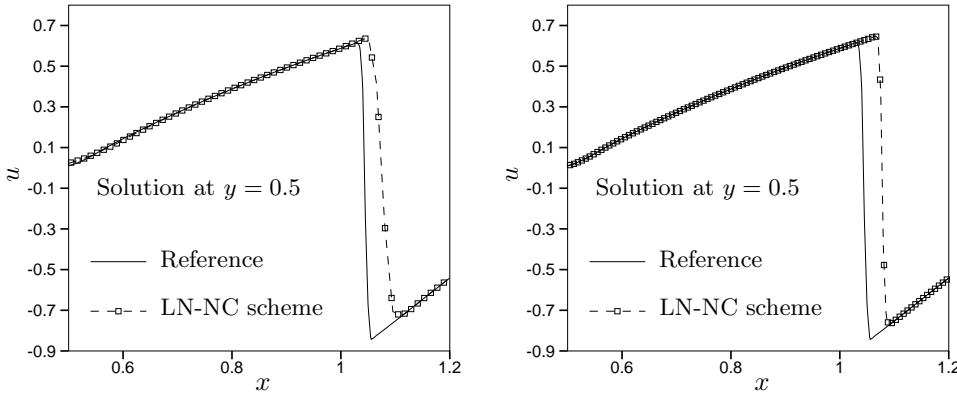


Figure 6.3: Nonlinear \mathcal{CL} with *exponential flux*: conservation error. Solution at $y = 0.5$ obtained with the LN-NC scheme on the coarse mesh (left) and on the fine mesh (right)

To understand the reason of this inconsistency, we have to recall the characterization of solutions to nonlinear conservation laws given in §2.2.2. There, we considered weak solutions which are piecewise smooth and separated by a discontinuity across which the local Rankine-Hugoniot jump conditions (2.18) are respected. As remarked in §2.2.2, these conditions enforce across a discontinuity the conservation of the unknown. The solution of our motivational example fits perfectly into this characterization. Since the problems seems to arise in correspondence of the discontinuity, we have a good hint that the non-conservative LN scheme behaves differently from the scheme used to compute the reference solution, with respect to the approximation of the Rankine-Hugoniot conditions. In particular, we know that across the discontinuity, the relevant form of the equation is the conservation law form. Hence, a proper definition of the element residual would take this into account, in formula

$$\phi^{\mathcal{CL}} = \int_E \nabla \cdot \mathcal{F}(u_h) \, dx \, dy = \oint_{\partial E} \mathcal{F}(u_h) \cdot \hat{n} \, dl$$

where the super-script \mathcal{CL} indicates that the residual is computed integrating the conservation law form (or divergence form) of the equation. What we have done is to linearize the problem and use the linearity of u_h , as follows

$$\begin{aligned} \phi^{\text{NC}} &= \int_E \nabla \cdot \mathcal{F}(u_h) \, dx \, dy = \int_E \tilde{a}(u_h) \cdot \nabla u_h \, dx \, dy = \\ &\quad \left(\int_E \tilde{a}(u_h) \, dx \, dy \right) \cdot \nabla u_h|_E \approx \overbrace{\frac{|E|}{3} \sum_{j \in E} \tilde{a}(u_j)}^{\text{inexact !}} \nabla u_h|_E \neq \phi^{\mathcal{CL}} \end{aligned}$$

We see that, since the linearization we introduced for \tilde{a} is inexact, we spoil the equivalence between the divergence form of the equation and its quasi-linear form. Hence, we do not approximate correctly the Rankine-Hugoniot conditions across the shock, thus obtaining a wrong result. The expression *non-conservative* scheme will then be used to refer to a scheme which is not capable of reproducing conditions (2.18) at the discrete level, rather than to a scheme built on the quasi-linear form of the equation.

6.2 Conservative \mathcal{RD} formulations: \mathcal{CRD} and \mathcal{QRD}

We start with the following definition

Definition 6.2.1 (Conservative \mathcal{RD} scheme). A \mathcal{RD} scheme is conservative if there exist a continuous approximation of the flux \mathcal{F}_h such that

$$\phi^h = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n} \, dl \tag{6.3}$$

This definition ensures that, across a stationary discontinuity, the steady limit of (2.18) is consistently reproduced by the element residual. In [5, 6] it is proved that under

assumptions of continuity of the split residuals and of the flux \mathcal{F}_h , conservative \mathcal{RD} schemes respect a Lax-Wendroff theorem. The non-conservative LN scheme does not verify definition 6.2.1, due to the introduction of the approximate Jacobian linearization. Hence, it does not respect the Lax-Wendroff theorem and, as a consequence, it converges to a wrong solution, as seen from the example.

Unfortunately, we have arrived to a problem of *incompatibility* between the use of the \mathcal{CL} form of the equation, needed to guarantee the approximation of the correct weak solution, and the use the flux Jacobians, needed in the definition of the k_j parameters used in the distribution of the residual. This incompatibility is particularly important if we want to make use of the limited N scheme. In this case, in fact, on one hand we need to be able to use the N scheme which, as originally formulated (see equations (5.43) and (5.11) or (5.7)), is largely based on the use of the quasi-linear form of the problem. It might be argued that a conservative nonlinear scheme could be constructed starting from a non-conservative N scheme. This would be the case if the element residual ϕ^h was computed in a conservative way, *e.g.* using contour integration of the fluxes, while the N scheme local nodal residuals were computed using an inexact linearization of the quasi-linear form. This case, however, would fall into the analysis of §5.5.2.1 and §5.5.2.2. In particular, the limiting procedure would not be well posed. Hence, we need to satisfy the conservation requirements and the conditions for the well-posedness of the limiting procedure (equations (5.67) and (5.68)). To cope with this issue, two solutions have been proposed which we present in the following subsections.

6.2.1 The QRD formulation

The analysis of the non-conservative scheme used to solve our motivational example leads to an idea which has been used in [4] to construct a class of \mathcal{FS} schemes based on the use of the quasi-linear form of the equation but guaranteeing a correct approximation of weak solutions of the nonlinear problem. To describe this approach, using the local regularity of $\mathcal{F}(u_h)$, we start by rewriting the element residual ϕ^h as

$$\begin{aligned} \phi^h &= \oint_{\partial E} \mathcal{F}(u_h) \cdot \hat{n} \, dl = \int_E \nabla \cdot \mathcal{F}(u_h) \, dx \, dy = \int_E \tilde{a}(u_h) \cdot \nabla u_h \, dx \, dy = \\ &= \left(\int_E \tilde{a}(u_h) \, dx \, dy \right) \cdot \nabla u_h|_E = |E| \tilde{a} \cdot \nabla u_h|_E = \sum_{j \in E} \tilde{k}_j u_j \end{aligned}$$

where now, we suppose that

$$\tilde{a} = \frac{1}{|E|} \int_E \tilde{a}(u_h) \, dx \, dy$$

is computed exactly. These schemes fit into the framework of definition 6.2.1 with $\mathcal{F}_h = \mathcal{F}(u_h)$, and u_h piecewise linear, as in (3.7). This means that conservative \mathcal{RD} schemes can be built if an *exact mean-value linearization* of the Jacobian is used.

The derivation of such a linearization can be difficult and in the case of a system of conservation laws almost impossible. This has motivated the authors of [4] to introduce an *approximate mean-value linearization* obtained with the Gaussian integration

$$\bar{a} = |E| \sum_{l=1}^{NQ} \omega_l \bar{a}(u(x_l, y_l)), \quad (x_l, y_l) \in E \quad (6.4)$$

where ω_l is the quadrature weight corresponding to the l -th Gaussian point (x_l, y_l) . As a consequence, the local residual can be expressed as

$$\phi^h = \sum_{j \in E} \bar{k}_j u_j = \int_E \nabla \cdot \mathcal{F}(u_h) dx dy + R_{NQ} \quad (6.5)$$

with R_{NQ} the *conservation error* due to the approximate integration of $\bar{a}(u_h)$. The properties of the Gaussian integration, namely the behavior of the quadrature error, allows the authors of [4] to prove that

- (a) provided that the number of quadrature points NQ is large enough, the conservation error due to the approximate integration is strictly smaller than the discretization error of the schemes;
- (b) Lax-Wendroff theorem: provided that the number of quadrature points NQ is large enough and under some continuity assumptions on the split residuals ϕ_i , \mathcal{RD} schemes based on the approximate Gaussian quadrature of the quasi-linear form of the problem converge to the correct weak solutions.

For brevity, we will refer to the schemes obtained using this approach as to the \mathcal{QRD} schemes, since conservation and convergence to the correct weak solutions is guaranteed by the accurate Quadrature of the quasi-linear form. Computationally speaking, \mathcal{QRD} schemes can be quite expensive, due to the evaluation of the Jacobians in NQ Gaussian points and to the fact that NQ has to be large enough to guarantee the quadrature error to be small enough [4]. However, due to the equivalence between the (approximate) linearized quasi-linear form and the conservation law form of the problem, the schemes presented in chapter 5 can be immediately used for the solution of nonlinear \mathcal{CL} s, the main problem being solved with the introduction of the *conservative* mean-value Jacobians. Moreover, the analysis of the discretization is formally identical to the linear case. In particular, even if we do not reserve a detailed discussion to this aspect, we stress that all the LED schemes presented in chapter 5, namely the upwind $\mathcal{FV} - \mathcal{RD}$ scheme, the R_v scheme and the N scheme, are still LED in the nonlinear case and they still respect the same positivity criteria and related stability bounds. Concerning their dissipation properties, as seen in chapter 2 the energy stability has to be replaced by the entropy stability, which is more appropriate in the nonlinear case. This topic will be covered later with some detail. Lastly, we note that the \mathcal{QRD} formulation solves at once not only the conservation problem but also the problem of guaranteeing the well-posedness of the limiting procedure, since once we locally linearize the equation, we can apply the N scheme as formulated in (5.43) which now respects (5.68) by construction. In particular, the reference solution of figure 6.1 has been computed with the limited N scheme constructed by integrating (6.4) with a 4 points formula.

6.2.2 The CRD formulation

While being mathematically sound, the \mathcal{QRD} approach has in its cost a major weakness which becomes even more important when dealing with systems of \mathcal{CL} s, as we will see later. A simpler solution has been proposed in [50], and it is the one used in this thesis. The first element of the construction is the definition of the element residual. Given a continuous approximation of the flux \mathcal{F}_h , we compute ϕ^h as

$$\phi^h = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n} \, dl = \sum_{l_j=1}^3 \mathcal{F}^{l_j} \cdot \vec{n}_{l_j} \quad \text{with} \quad \mathcal{F}^{l_j} = \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}_h(x_p, y_p), \quad (x_p, y_p) \in l_j \quad (6.6)$$

where l_1 , l_2 and l_3 are the edges of E , \vec{n}_{l_j} is the exterior normal to l_j , scaled by the length of the edge, and ω_p is the weight of the p -th quadrature point on l_j . As before, the computation of the residual is based on a quadrature formula, however, differently from the \mathcal{QRD} case, given a quadrature formula which is at least exact for a linear flux \mathcal{F}_h , definition (6.6) satisfies by construction (6.3), more accurate quadrature corresponding to a more accurate *reconstruction* of \mathcal{F}_h on the edges of the element. For example, if $\mathcal{F}_h = \mathcal{F}(u_h)$, with u_h linear, the application of Simpson's quadrature formula on l_j corresponds to a parabolic representation of \mathcal{F} using its values in the limiting nodes of l_j and in the midpoint. Not to be misled by this observation, note that this does not mean that we can increase the accuracy of the schemes by selecting a more accurate formula, since the values of \mathcal{F} in internal points of l_j are evaluated using the linear variation of u_h between the extremities of the edge, which ultimately limits the accuracy we can achieve to second-order. In particular, recalling the analysis of §4.4.1 (equation (4.48)), second-order conservative schemes can be obtained already evaluating (6.6) with the trapezium rule. As in [50], we refer to schemes based on this definition of the residual as to \mathcal{CRD} schemes, since conservation is guaranteed by the use of Contour integration for the evaluation of ϕ^h . Definition (6.6) of the residual ensures conservation, however we need to specify how the flux Jacobians can be used to distribute ϕ^h , to complete the \mathcal{CRD} formulation of the schemes.

6.2.2.1 Linearity preserving schemes

The case of \mathcal{LP} schemes is quite simple, since these schemes are defined by

$$\phi_i = \beta_i \phi^h$$

with β_i uniformly bounded and respecting the consistency condition

$$\sum_{j \in E} \beta_j = 1$$

by construction. The dependence of the distribution coefficients on the k_j parameters does not alter any of these two properties (boundedness and consistency). In particular, we can use for the computation of the β_i s the parameters

$$k_j = \frac{\vec{a}(u_E) \cdot \vec{n}_j}{2} \quad (6.7)$$

with u_E an arbitrary average of u_h over E . For example, one easily checks that consistency is guaranteed for the LDA scheme (5.39) and for the SUPG-like schemes (5.27). To distinguish between this case and the \mathcal{QRD} case, we will denote by $\bar{\beta}_i$ the distribution coefficients evaluated making use of the \bar{k}_j parameters based on the approximate quadrature (6.4), while keeping the notation β_i for the ones making use of (6.7). Obviously, in general

$$\beta_i \neq \bar{\beta}_i$$

6.2.2.2 The N scheme

The basic idea that led the authors of [50] to the formulation of the \mathcal{CRD} approach is contained in equation (5.47) and in the analysis of §5.4.3. We will recall here this idea and also show some of its implications and relations with previously published techniques. The basic observation done in [50] is that the conditions

$$\begin{cases} \phi_i^N &= k_i^+(u_i - u_{in}) \\ \phi^h &= \sum_{j \in E} \phi_j^N \end{cases} \quad (6.8)$$

uniquely define u_{in} . In particular, if k_j is computed using an arbitrary average u_E as in (6.7), and ϕ^h is given by (6.6), the definition

$$u_{in} = \left(\sum_{j \in E} k_j^+ \right)^{-1} \left(\sum_{j \in E} k_j^+ u_j - \phi^h \right) = N \left(\sum_{j \in E} k_j^+ u_j - \phi^h \right) \quad (6.9)$$

gives the unique state u_{in} guaranteeing that a scheme formally identical to the N scheme (5.43) satisfies the \mathcal{RD} consistency condition with respect to the conservative definition (6.6) of the residual. The analysis of §5.4.3, however, enables to say something more. In particular, in this thesis, we refer to the \mathcal{CRD} N scheme, as to the scheme defined by the local nodal residuals

$$\phi_i^{N-\mathcal{CRD}} = \phi_i^{\text{LDA}-\mathcal{CRD}} + d_i^{N-\mathcal{CRD}} \quad (6.10)$$

where the dissipation terms $d_i^{N-\mathcal{CRD}}$ are given by

$$d_i^{N-\mathcal{CRD}} = \sum_{j \in E} k_i^+ N k_j^+ (u_i - u_j) = k_i^+ (u_i - u_{out}) \quad (6.11)$$

with

$$\sum_{j \in E} d_j^{N-\mathcal{CRD}} = 0 \quad (6.12)$$

Clearly, the satisfaction of the \mathcal{RD} consistency condition for the LDA scheme being guaranteed independently on how we evaluate the k_j parameters, and condition (6.12) also being always respected, formulating the N scheme as the LDA plus dissipation gives a natural way of extending the scheme to situations in which the residual ϕ^h is defined in a way as general as possible. A second interpretation of this formulation of the N

scheme is the one already given in [44], linking the \mathcal{CRD} approach to the techniques proposed in [92, 48]. In the last references the authors propose an approach allowing to use the N scheme without having to resort to exact mean-value linearizations of the flux Jacobians. The idea is to define the local nodal residuals as

$$\phi_i^{\text{N-C}_{corr}} = k_i^+(u_i - u_{in}) + \beta_i \epsilon_{\mathcal{F}}, \quad u_{in} = - \sum_{j \in E} k_j^- N u_j$$

where the first term represents the non-conservative scheme obtained by blindly applying (5.43) with the k_j parameters computed using an inexact linearization, as in (6.7). The second term, instead, is a *conservative correction* proportional to the error

$$\epsilon_{\mathcal{F}} = \phi^h - \sum_{j \in T} k_j^+(u_j - u_{in}) = \phi^h - \sum_{j \in T} k_j u_j$$

with ϕ^h given for example by (6.6). In this formulation, the distribution coefficients β_i used to split the conservation error remain somehow undefined. In [92, 48], several choices have been tested. However, as observed in [44], if we use $\beta_i = \beta_i^{\text{LDA}}$ to distribute the error, then, using the definition of β_i^{LDA} (equation (5.39)), and relation (5.11):

$$\begin{aligned} \phi_i^{\text{N-C}_{corr}} &= k_i^+(u_i - u_{in}) + \beta_i^{\text{LDA}} \left(\phi^h - \sum_{j \in T} k_j u_j \right) = \\ &= \beta_i^{\text{LDA}} \phi^h + k_i^+(u_i - u_{in}) - \beta_i^{\text{LDA}} \sum_{j \in T} k_j u_j = \\ &= \phi_i^{\text{LDA-CRD}} + k_i^+(u_i - u_{in}) - k_i^+ N \left(\sum_{j \in T} k_j^+ \right) (u_{out} - u_{in}) = \\ &= \phi_i^{\text{LDA-CRD}} + k_i^+(u_i - u_{in}) - k_i^+(u_{out} - u_{in}) = \phi_i^{\text{LDA-CRD}} + d_i^{\text{N-CRD}} = \phi_i^{\text{N-CRD}} \end{aligned}$$

This shows the equivalence of the conservative correction technique of [92, 48] with the \mathcal{CRD} formulation of the N scheme. In particular that the distribution of the conservation error is uniquely determined by the constraint that the resulting scheme is of the form (6.10) (or equivalently (5.43)).

Formulating the N scheme as the LDA scheme plus dissipation gives a flexible way of changing the definition of the residual ϕ^h while keeping fixed the dissipation that the scheme adds to the LDA. In §5.4.4 this has been used to design an extension of the N scheme to advection problems with solution independent source terms. In that case, we have still been able to prove the L^∞ stability of the resulting scheme. Unfortunately, we cannot do the same for the \mathcal{CRD} N scheme, for which, instead, we have a negative result. To present this result, we assume that

Assumption 6.2.2 (Bridge between \mathcal{QRD} and \mathcal{CRD} schemes). *Given a NC-points line quadrature formula used to evaluate (6.6), it is possible to find a NQ-surface quadrature rule to be used in (6.4), such that the equivalence*

$$\sum_{l_j=1}^3 \sum_{p=1}^{NC} \omega_p \mathcal{F}(u_{p,l_j}) \cdot \vec{n}_{l_j} = |E| \sum_{l=1}^{NQ} \omega_l \vec{a}(u_l) \cdot \nabla u_h|_E = \sum_{j \in E} \bar{k}_j u_j \quad (6.13)$$

holds up to the smallest between the quadrature error in (6.6), the error in (6.5) and the discretization error of the schemes.

This equivalence sets a connection between the \mathcal{QRD} and the \mathcal{CRD} approaches. In particular, \mathcal{QRD} schemes are the only \mathcal{CRD} schemes for which the average speed used to compute the \bar{k}_j parameters coincides with the conservative mean-value linearized speed. Using this equivalence, we can now prove the following negative result.

Proposition 6.2.3 (\mathcal{CRD} N scheme and sub-element LED). *The \mathcal{CRD} N scheme (6.10) cannot be proved to respect the sub-element LED condition. In particular, the scheme is prone to the violation of this condition in multiple-target elements.*

Proof. The proof is obtained by exploiting the equivalence between the residual evaluated with the contour integral and the one obtained with the mean-value linearization of the Jacobians. In particular, this equivalence allows to recast the local nodal residuals of the \mathcal{CRD} N scheme as

$$\phi_i^{\text{N-}\mathcal{CRD}} = k_i^+ N \sum_{j \in E} \bar{k}_j u_j + \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N k_j^+ (u_i - u_j)$$

with \bar{k}_j evaluated using the conservative mean-value Jacobians. Since the \bar{k}_j s sum up to zero over an element (see (3.17)), we can rewrite the expression as

$$\begin{aligned} \phi_i^{\text{N-}\mathcal{CRD}} &= \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N \bar{k}_j (u_j - u_i) + \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N k_j^+ (u_i - u_j) = \\ &= \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N \bar{k}_j^+ (u_j - u_i) + \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N \bar{k}_j^- (u_j - u_i) + \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N k_j^+ (u_i - u_j) = \\ &= \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N (k_j^+ - \bar{k}_j^+) (u_i - u_j) - \sum_{\substack{j \in E \\ j \neq i}} k_i^+ N \bar{k}_j^- (u_i - u_j) \end{aligned}$$

Last expression shows that whenever $k_j = \bar{k}_j$, as in the linear case or with \mathcal{QRD} schemes, the scheme reduces to its standard expression, which verifies the sub-element LED condition with $c_{ij}^E = -\bar{k}_i^+ N \bar{k}_j^- \geq 0$. In general, however, we have

$$c_{ij}^E = k_i^+ N (k_j^+ - \bar{k}_j^+) - \overbrace{k_i^+ N \bar{k}_j^-}^{\geq 0}$$

Since the sign of the first term on the right hand side is unknown, we cannot prove the sub-element LED condition. Consider now the multiple-target situation $k_i, \bar{k}_i, \bar{k}_j > 0$ for a node $j \neq i$. In this case we have

$$c_{ij}^E = k_i^+ N (k_j^+ - \bar{k}_j^+)$$

where the beneficial effect of the second term has disappeared. The sign of c_{ij}^E could be either positive or negative, depending on the local structure of the solution and on the average used for the evaluation of k_j . Hence, the scheme is particularly prone to the violation of the local LED condition in multiple-target elements. \square

From the theoretical point of view, this result is quite disappointing, since it seems to spoil our hopes of constructing a non-oscillatory approximation of nonlinear \mathcal{CL} s. In particular, while on twodimensional triangular grids the non-LED character of the \mathcal{CRD} N scheme could be limited to a few 2-target elements, in three space dimensions things could get much worse due to the presence of a larger number of 2-target tetrahedra and of 3-target elements which are not present in 2D. Perhaps surprisingly, these effects have never been observed in any numerical calculation both in two and three space dimensions, for scalar problems and also for systems of conservation laws [50, 44]. Similarly, the extension of this scheme to meshes composed of quadrilaterals [134, 63] has shown quite a non-oscillatory character of the numerical solutions, both for scalar problems and for systems. We believe that the monotone resolution of discontinuities observed in practice is due partly to a compensation of the local violation of the LED condition predicted by proposition 6.2.3 when assembling the contributions of all the elements surrounding a node, and more importantly to the dissipative character of the scheme, related to the form of the $d_i^{\text{N-}\mathcal{CRD}}$ terms (6.11). This last effect is probably enough to dissipate local new extrema eventually appearing in the numerical solution. The results of this thesis contribute to confirm the robustness of the \mathcal{CRD} N scheme.

More importantly, the scheme verifies by construction the sufficient condition for the well-posedness of the limiting (equation (5.68)). Hence, we can apply mapping (5.65) to obtain a \mathcal{LP} nonlinear scheme. In this thesis, we give extensive numerical evidence of the reliability, robustness and accuracy of this \mathcal{CRD} limited N scheme.

6.2.2.3 Rusanov's scheme

The extension of the Rv scheme to the \mathcal{CRD} framework is quite immediate, due to the fact that this scheme is formulated as the central scheme plus an isotropic dissipation term. Hence we have

$$\phi_i^{\text{Rv-}\mathcal{CRD}} = \frac{1}{3}\phi^h + d_i^{\text{Rv-}\mathcal{CRD}}, \quad d_i^{\text{Rv-}\mathcal{CRD}} = \frac{1}{3}\alpha \sum_{\substack{j \in E \\ j \neq i}} (u_i - u_j) \quad (6.14)$$

Differently from the N scheme, due to the isotropic character of $d_i^{\text{Rv-}\mathcal{CRD}}$ the Rv scheme respects the sub-element LED criterion if α is large enough as shown by

$$\phi_i^{\text{Rv-}\mathcal{CRD}} = \sum_{\substack{j \in E \\ j \neq i}} c_{ij}^E (u_i - u_j), \quad c_{ij}^E = \frac{1}{3}(\alpha - \bar{k}_j)$$

However, we *will not* use the Rv scheme for the construction of a limited nonlinear scheme. The reason of this choice is basically the lack of a sufficient understanding of the properties of the limited schemes. As already remarked in §5.5.2.2, the dissipation properties of these nonlinear schemes are very unclear at the moment. The available technology works best with schemes with a pronounced upwind character. It would be however of great practical importance to be able to extend this technology to simpler first-order schemes, such as (6.14).

6.3 Conservative residual distribution and entropy

The stability of schemes approximating nonlinear \mathcal{CL} s is characterized by the satisfaction of a discrete analog of the entropy inequality (2.20). As we saw in §2.2.3, in the continuous case this inequality corresponds to the presence of a vanishing dissipative mechanism which uniquely determines the exact weak solution of the problem. Similarly, in the discrete case the satisfaction of this stability criterion is related to the presence of a dissipation mechanism acting on the discrete unknown. The analysis of this section partly follows [5, 6, 4]. In particular, let us consider a convex entropy pair $(\mathcal{H}(u), \mathcal{G}(u))$ ¹ supplementing equation (6.1), such that weak exact solutions of the problem satisfy in a weak sense

$$\frac{\partial \mathcal{H}(u)}{\partial t} + \nabla \cdot \mathcal{G}(u) \leq 0 \quad (6.15)$$

As anticipated in §3.2.2, in order to study the entropy stability of \mathcal{RD} discretizations, we introduce the entropy variable v defined as

$$v = \frac{d\mathcal{H}(u)}{du}$$

The analysis of the stability of the schemes will be performed by assuming that v is used as a primary variable (see §3.2). In particular, given the initial data $u_0(x, y)$, we will analyze the \mathcal{RD} prototype written in terms of the entropy variable v :

$$|S_i| \frac{du(v_i)}{dt} + \sum_{E \in \mathcal{D}_i} \phi_i(v_h), \quad (6.16)$$

with v_h piecewise linear as in (3.11) and with

$$\sum_{j \in E} \phi_j(v_h) = \int_E \nabla \cdot \mathcal{F}(v_h) \, dx \, dy \quad (6.17)$$

with $\mathcal{F}(v_h) = \mathcal{F}(u(v_h))$. Also note that, as anticipated in §3.3, in the following pages we keep the notation $\vec{a}(v) = \vec{a}(u(v))$ and k_j for the flux Jacobian and the upwind parameters, which are now computed using flux derivatives with respect to v .

The semi-discrete entropy balance for \mathcal{RD} schemes in entropy variable is obtained by multiplying by v_i the semi-discrete \mathcal{RD} evolution equation for node i and summing over all the nodes of the mesh:

$$\sum_{i \in \mathcal{T}_h} v_i |S_i| \frac{du(v_i)}{dt} + \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} v_i \phi_i(v_h) = 0$$

Due to the definition of v , last expression is a second-order approximation of the global entropy evolution equation

$$\int_{\Omega} \frac{d\mathcal{H}(v_h)}{dt} + \sum_{E \in \mathcal{T}_h} \sum_{j \in E} v_j \phi_j(v_h)$$

¹see the definition given in §2.2.3, equation (2.19)

Introducing the *local entropy production* of a scheme

$$\Phi_{\mathcal{H}}^E(v_h) = \sum_{j \in E} v_j \phi_j(v_h) \quad (6.18)$$

a general entropy balance for \mathcal{RD} scheme is obtained:

$$\int_{\Omega} \frac{d\mathcal{H}(v_h)}{dt} + \sum_{E \in \mathcal{T}_h} \Phi_{\mathcal{H}}^E(v_h) = 0 \quad (6.19)$$

Last equation gives a criterion to determine whether a scheme is stable or not.

Definition 6.3.1 (Entropy conservative \mathcal{RD} scheme). *A \mathcal{RD} scheme in entropy variable is entropy conservative if there is a continuous consistent approximation of the entropy flux \mathcal{G}_h such that*

$$\Phi_{\mathcal{H}}^E(v_h) = \int_E \nabla \cdot \mathcal{G}_h \, dx \, dy = \oint_{\partial E} \mathcal{G}_h \cdot \hat{n} \, dl \quad (6.20)$$

The Lax-Wendroff theorem recalled in the beginning of chapter 4 guarantees that, under the hypotheses of continuity of \mathcal{G}_h and of the local nodal residuals $\phi_i(v_h)$, the discrete solution obtained with an entropy conservative scheme converges to a weak-solution of the conservation law obtained taking the equal sign in (6.15) [5, 6]. Hence, entropy conservative schemes satisfy (in a weak sense) an *entropy equality*. This means that entropy conservative schemes are basically non-dissipative.

Definition 6.3.2 (Entropy stable \mathcal{RD} scheme). *A \mathcal{RD} scheme in entropy variable is entropy stable if there is a continuous consistent approximation of the entropy flux \mathcal{G}_h such that*

$$\Phi_{\mathcal{H}}^E(v_h) = \int_E \nabla \cdot \mathcal{G}_h \, dx \, dy + \epsilon_{\mathcal{H}}^E = \oint_{\partial E} \mathcal{G}_h \cdot \hat{n} \, dl + \epsilon_{\mathcal{H}}^E, \quad \epsilon_{\mathcal{H}}^E \geq 0 \quad (6.21)$$

The presence of the *numerical dissipation* $\epsilon_{\mathcal{H}}^E$ and the Lax-Wendroff theorem guarantee that the discrete solution obtained with entropy stable schemes also satisfy the *entropy inequality* (6.15) in a weak sense [5, 6]. Entropy stable schemes are the ones introducing a *vanishing* dissipative mechanism into the discrete equations. As we will see, the analysis of the entropy stability properties of \mathcal{RD} schemes is very similar to the energy stability analysis performed in chapter 5. However, differently from \mathcal{FE} schemes, it is hard to give for \mathcal{RD} scheme a real stability proof, while it is generally possible to show some *weaker* consistency with the entropy equality or inequality. This consistency not constituting a real stability proof, it nevertheless shows the presence of a dissipative mechanism acting on the discrete solution. In the following pages, we will only consider linear schemes, the analysis of the nonlinear schemes being substantially similar to, and as approximate as, their energy stability analysis reported in §5.5. We underline once again that, theoretically speaking, the understanding of the stability properties of nonlinear \mathcal{RD} discretizations remains one of the most challenging open issues.

6.3.1 Entropy stability, central \mathcal{RD} schemes and \mathcal{FE}

6.3.1.1 Galerkin \mathcal{FE} and centered \mathcal{FS} scheme

The case of a pure central scheme well describes the basic difference between the \mathcal{FE} discretization and the \mathcal{FS} one. In particular, let us consider the local nodal residual of the Galerkin scheme

$$\phi_i^G = \int_E \psi_i \nabla \cdot \mathcal{F}(v_h) \, dx \, dy$$

with ψ_i the basis function of node i (see §3.2.1). The local entropy production of the scheme is

$$\Phi_{\mathcal{H}}^G = \sum_{j \in E} v_j \psi_j^G = \sum_{j \in E} \int_E \psi_j v_j \nabla \cdot \mathcal{F}(v_h) \, dx \, dy$$

We immediately recognize that, due to the definition of v

$$\Phi_{\mathcal{H}}^G = \int_E v_h \nabla \cdot \mathcal{F}(v_h) \, dx \, dy = \int_E \nabla \cdot \mathcal{G}(v_h) \, dx \, dy$$

Hence, with the obvious choice $\mathcal{G}_h = \mathcal{G}(v_h)$ the Galerkin \mathcal{FE} scheme is entropy conservative. Consider now the case of the central \mathcal{RD} scheme

$$\phi_i^C = \frac{1}{3} \int_E \nabla \cdot \mathcal{F}_h \, dx \, dy$$

One possibility we have is to use the \mathcal{CRD} formulation of the scheme based on a piecewise continuous linear flux \mathcal{F}_h , in which case $\nabla \cdot \mathcal{F}_h$ is constant over E , giving for the local entropy production

$$\Phi_{\mathcal{H}}^C = \int_E \sum_{j \in E} \frac{v_j}{3} \nabla \cdot \mathcal{F}_h \, dx \, dy = \int_E v_h \nabla \cdot \mathcal{F}_h \, dx \, dy$$

Even though formally similar to $\Phi_{\mathcal{H}}^G$, this expression does not really define a continuous approximation of the entropy flux \mathcal{G}_h such that the conservative definition (6.20) is respected, due to the poor approximation of \mathcal{F} . However, using the equivalence of assumption 6.2.2, we can write (see also [4])

$$\Phi_{\mathcal{H}}^C = \int_E v_h \bar{a} \cdot \nabla v_h \, dx \, dy \xrightarrow{h \rightarrow 0} \int_E v \bar{a}(v) \cdot \nabla v \, dx \, dy = \int_E \nabla \cdot \mathcal{G}(v) \, dx \, dy$$

where \bar{a} is the conservative mean-value linearized flux Jacobian. Hence, even though not exactly entropy conservative, the central scheme gives an approximation consistent with the entropy equality. For this reason, we introduce the following definitions.

Definition 6.3.3 (Entropy consistent scheme). *A \mathcal{RD} scheme is entropy consistent if it is possible to show that*

$$\Phi_{\mathcal{H}}^E \xrightarrow{h \rightarrow 0} \int_E \nabla \cdot \mathcal{G}(v) \, dx \, dy$$

Definition 6.3.4 (Entropy dissipative scheme). A \mathcal{RD} scheme is entropy dissipative if it is possible to show that

$$\Phi_{\mathcal{H}}^E = \Phi_{\mathcal{G}}^E + \epsilon_{\mathcal{H}}^E \quad \text{with} \quad \Phi_{\mathcal{G}}^E \xrightarrow{h \rightarrow 0} \int_E \nabla \cdot \mathcal{G}(v) \, dx \, dy$$

and with $\epsilon_{\mathcal{H}}^E \geq 0$.

Clearly, the central \mathcal{RD} scheme is entropy consistent and non-dissipative. Schemes which are more dissipative than the central scheme are entropy dissipative. Note that the application of the dominated convergence theorem to an entropy dissipative scheme implies that, if the scheme is convergent, then the solution obtained in the limit $h \rightarrow 0$ respects an entropy inequality (see [4] for more).

6.3.1.2 Streamline dissipation

An entropy stable scheme is the SUPG \mathcal{FE} scheme [166, 72, 94, 97, 98, 96, 102, 103]

$$\phi_i^{\text{SUPG}} = \phi_i^{\text{G}} + \int_E \tau (\vec{a}(v_h) \cdot \nabla \psi_i) (\vec{a}(v_h) \cdot \nabla v_h) \, dx \, dy$$

In fact, proceeding as before, we have for the SUPG scheme

$$\Phi_{\mathcal{H}}^{\text{SUPG}} = \Phi_{\mathcal{H}}^{\text{G}} + \int_E \tau \left(\sum_{j \in E} \vec{a}(v_h) \cdot \nabla \psi_j v_j \right) (\vec{a}(v_h) \cdot \nabla v_h) \, dx \, dy$$

leading to

$$\Phi_{\mathcal{H}}^{\text{SUPG}} = \Phi_{\mathcal{H}}^{\text{G}} + \int_E \tau (\vec{a}(v_h) \cdot \nabla v_h)^2 \, dx \, dy = \int_E \nabla \cdot \mathcal{G}(v_h) \, dx \, dy + \epsilon_{\mathcal{H}}^{\text{SUPG}}$$

which shows the stability of the scheme. In the case of the linearity preserving \mathcal{RD} PG scheme with distribution coefficients (5.27), we have to make a distinction between the \mathcal{QRD} and the \mathcal{CRD} formulations of the scheme. In this first case we have

$$\phi_i^{\text{PG-QRD}} = \phi_i^{\text{C}} + \tau \frac{\bar{k}_i}{2|E|} \sum_{j \in E} \bar{k}_j u_j = \phi_i^{\text{C}} + \int_E \tau (\vec{a} \cdot \nabla \psi_i) (\vec{a} \cdot \nabla v_h) \, dx \, dy$$

giving the local entropy production

$$\Phi_{\mathcal{H}}^{\text{PG-QRD}} = \Phi_{\mathcal{H}}^{\text{C}} + \int_E \tau (\vec{a} \cdot \nabla v_h)^2 \, dx \, dy = \Phi_{\mathcal{H}}^{\text{C}} + \epsilon_{\mathcal{H}}^{\text{PG}}$$

However, for the \mathcal{CRD} scheme defined by

$$\phi_i^{\text{PG-CRD}} = \phi_i^{\text{C}} + \tau \frac{k_i}{2|E|} \sum_{j \in E} \bar{k}_j u_j = \phi_i^{\text{C}} + \int_E \tau (\vec{a}(v_E) \cdot \nabla \psi_i) (\vec{a} \cdot \nabla v_h) \, dx \, dy$$

we have the local entropy production

$$\Phi_{\mathcal{H}}^{\text{PG-}\mathcal{CRD}} = \Phi_{\mathcal{H}}^{\text{C}} + \overbrace{\int_E \tau (\bar{a}(v_E) \cdot \nabla v_h) (\bar{a} \cdot \nabla v_h) \, dx \, dy}^{\geq \text{ or } \leq 0 \, ?}$$

where, strictly speaking, due to the introduction of the inexact linearization of the Jacobian $\bar{a}(v_E)$, the streamline dissipation term is not guaranteed to locally dissipate anymore, unless some hypotheses on the products $k_j \bar{k}_j$ are introduced. This means that, while in its \mathcal{QRD} formulation the PG scheme is entropy dissipative, the \mathcal{CRD} formulation does not guarantee the preservation of this property, unless the quadratic form associated to the \mathcal{CRD} streamline dissipation (see equation (5.29))

$$Q^{\text{PG}} = \frac{\tau}{4|E|} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}^T \begin{bmatrix} 2k_1 \bar{k}_1 & k_1 \bar{k}_2 + k_2 \bar{k}_1 & k_1 \bar{k}_3 + k_3 \bar{k}_1 \\ k_1 \bar{k}_2 + k_2 \bar{k}_1 & 2k_2 \bar{k}_2 & k_2 \bar{k}_3 + k_3 \bar{k}_2 \\ k_1 \bar{k}_3 + k_3 \bar{k}_1 & k_2 \bar{k}_3 + k_3 \bar{k}_2 & 2k_3 \bar{k}_3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

is positive.

6.3.1.3 The Rv scheme

In entropy variables, the Rv scheme reads

$$\phi_i^{\text{Rv}} = \phi_i^{\text{C}} + \frac{\alpha}{3} \sum_{\substack{j \in E \\ j \neq i}} (v_i - v_j)$$

with

$$\alpha > \max_{j \in E} \bar{k}_j > 0$$

The local entropy production is easily shown to be

$$\Phi_{\mathcal{H}}^{\text{Rv}} = \Phi_{\mathcal{H}}^{\text{C}} + \epsilon_{\mathcal{H}}^{\text{Rv}}, \quad \epsilon_{\mathcal{H}}^{\text{Rv}} = \frac{1}{3} \begin{bmatrix} v_1 - v_2 \\ v_1 - v_3 \\ v_2 - v_3 \end{bmatrix}^T \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix} \begin{bmatrix} v_1 - v_2 \\ v_1 - v_3 \\ v_2 - v_3 \end{bmatrix} \geq 0$$

which proves the dissipative character of the scheme.

6.3.2 Entropy dissipation and \mathcal{MU} schemes

The analysis of the LDA and N scheme is much more delicate and less developed. Some results have been published in [3, 4, 9] a few of which will be recalled here. Things are even more difficult in the case of the \mathcal{CRD} schemes for which we cannot fully benefit from the equivalence between the quasi-linear form of the problem and its \mathcal{CL} form. However, in simple configurations some results can be shown, as the following.

Proposition 6.3.5 (Entropy dissipation of \mathcal{MU} schemes: the 1-target case). *\mathcal{QRD} \mathcal{MU} schemes are locally entropy dissipative in the 1-target case. This applies to \mathcal{CRD} schemes only if $k_j \bar{k}_j > 0 \forall j$*

Proof. To obtain the proof we recall that in a 1-target element all \mathcal{MU} schemes distribute the whole residual to the only downstream node. Suppose then that node 1 coincides with the outflow node, such that $\bar{k}_1 > 0, \bar{k}_2, \bar{k}_3 < 0$ and

$$\phi_1 = \phi^h, \quad \phi_2 = \phi_3 = 0$$

We rewrite the scheme as $\phi_i = \phi_i^C + (\phi_i - \phi_i^C)$ so that the local entropy production is

$$\Phi_{\mathcal{H}}^{\mathcal{MU}} = \Phi_{\mathcal{H}}^C + \epsilon_{\mathcal{H}}^{\mathcal{MU}}$$

where $\epsilon_{\mathcal{H}}^{\mathcal{MU}}$ is given by

$$\epsilon_{\mathcal{H}}^{\mathcal{MU}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}^T \widetilde{M} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad \widetilde{M} = \frac{1}{3} \begin{bmatrix} 2\bar{k}_1 & 2\bar{k}_2 & 2\bar{k}_3 \\ -\bar{k}_1 & -\bar{k}_2 & -\bar{k}_3 \\ -\bar{k}_1 & -\bar{k}_2 & -\bar{k}_3 \end{bmatrix}$$

Since we also have

$$\epsilon_{\mathcal{H}}^{\mathcal{MU}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}^T M \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad M = \frac{\widetilde{M} + \widetilde{M}^T}{2}$$

we can prove that the scheme is dissipative by studying the properties of M :

$$M = \frac{1}{6} \begin{bmatrix} 4\bar{k}_1 & 2\bar{k}_2 - \bar{k}_1 & 2\bar{k}_3 - \bar{k}_1 \\ 2\bar{k}_2 - \bar{k}_1 & -2\bar{k}_2 & -\bar{k}_2 - \bar{k}_3 \\ 2\bar{k}_3 - \bar{k}_1 & -\bar{k}_2 - \bar{k}_3 & -2\bar{k}_3 \end{bmatrix}$$

Since node 1 is downstream and nodes 2 and 3 are upstream, M has positive entries on the diagonal and negative off-diagonal terms. Moreover, since $\bar{k}_1 + \bar{k}_2 + \bar{k}_3 = 0$, one immediately checks that the row-sum of the elements of M is identically zero. Hence M is positive semi-definite. This means that $\epsilon_{\mathcal{H}}^{\mathcal{MU}} \geq 0$, which proves that the scheme is dissipative. Clearly, if $k_j \bar{k}_j > 0$, \mathcal{CRD} schemes will treat this element as 1-target, and the analysis also applies to this case. \square

6.3.2.1 The LDA scheme

It is evident from proposition 6.3.5 that multidimensional upwinding has a beneficial effect in terms of stabilization. However, the analysis of 2-target cases is not as easy as in the 1-target case. In particular, the trick of comparing the schemes to the centered distribution does not pay off. In its \mathcal{QRD} formulation, the LDA scheme can still benefit from the possibility of decomposing its entropy production in a term which acts in a

central fashion along a local direction plus some dissipation (see §5.4.1.1, equations (5.41) and (5.42)):

$$\Phi_{\mathcal{H}}^{\text{LDA-}\mathcal{QRD}} = \frac{\bar{v}_{out} + \bar{v}_{in}}{2} \left(\sum_{j \in E} \bar{k}_j^+ \right) (\bar{v}_{out} - \bar{v}_{in}) + \frac{1}{2} (\bar{v}_{out} - \bar{v}_{in}) \left(\sum_{j \in E} \bar{k}_j^+ \right) (\bar{v}_{out} - \bar{v}_{in})$$

where \bar{v}_{out} and \bar{v}_{in} denote the outflow and inflow states computed using the upwind parameters \bar{k}_j , based on the conservative mean-value linearized Jacobian. As done in the energy analysis of the scheme, we can introduce, locally in a 2-target element E , a coordinate direction ζ such that the local entropy production becomes

$$\Phi_{\mathcal{H}}^{\text{LDA-}\mathcal{QRD}} = \int_{\zeta_E} v_h \bar{a}^* \frac{\partial v_h}{\partial \zeta} d\zeta + \epsilon_{\mathcal{H}}^{\text{LDA-}\mathcal{QRD}}, \quad \bar{a}^* = \sum_{j \in E} \bar{k}_j^+ \quad (6.22)$$

with

$$\epsilon_{\mathcal{H}}^{\text{LDA-}\mathcal{QRD}} = \frac{1}{2} (\bar{v}_{out} - \bar{v}_{in}) \left(\sum_{j \in E} \bar{k}_j^+ \right) (\bar{v}_{out} - \bar{v}_{in}) \geq 0$$

As in the linear case, the integral in (6.22) is formally similar to the entropy production of a onedimensional central scheme acting on ζ_E . It is however unclear how this term can be recast in a form giving back the integral of the divergence of the entropy flux, when summing all the elemental contributions. However, equation (6.22) shows the dissipative effects of the \mathcal{MU} . As in the case of the PG scheme, when using the \mathcal{CRD} formulation of the scheme things become less clear, due to the introduction of the inexact linearization of the flux Jacobian. For completeness, we report hereafter the local production of entropy of the scheme:

$$\Phi_{\mathcal{H}}^{\text{LDA-}\mathcal{CRD}} = v_{out} \left(\sum_{j \in E} \bar{k}_j^+ \right) (\bar{v}_{out} - \bar{v}_{in}) \quad (6.23)$$

where the equivalence of assumption 6.2.2 has been used to obtain the right-hand side. The basic problem is now that the upwinding direction depends on the flux Jacobian. The use of the inexact linearization for the distribution introduces then a direction parallel to $\bar{a}(v_E)$, on which the states v_{out} and v_{in} lay, different from the direction determined by \bar{a} , on which the states \bar{v}_{out} and \bar{v}_{in} lay. This makes impossible the use of the 1D analogy discussed for the \mathcal{QRD} scheme.

6.3.2.2 The N scheme

The case of the N scheme is of course very interesting, since this scheme is the basis of all the constructions of high-order discretizations. Being \mathcal{MU} , the scheme respects proposition 6.3.5. However, in the 2-target case, rewriting the N scheme as a central distribution plus extra terms does not lead to an immediate proof of the dissipative character of the scheme. In the \mathcal{QRD} case, this can be shown in a slightly more elaborate way. Following [4, 9], we recall that the energy matrix operator of the N scheme can be written as

$$M^{\text{N-}\mathcal{QRD}} = \bar{D}^{\text{N-}\mathcal{QRD}} + \frac{1}{2} \begin{bmatrix} \bar{k}_1 & 0 & 0 \\ 0 & \bar{k}_2 & 0 \\ 0 & 0 & \bar{k}_3 \end{bmatrix}$$

with $\overline{D}^{N-\mathcal{QRD}}$ the equivalent entropy operator (see §5.4.2.1, equation (5.46))

$$\begin{aligned} \overline{D}^{N-\mathcal{QRD}} = & \frac{1}{2} \begin{bmatrix} \overline{k}_1 \\ \overline{k}_2 \\ \overline{k}_3 \end{bmatrix} \overline{N} \begin{bmatrix} \overline{k}_1 \\ \overline{k}_2 \\ \overline{k}_3 \end{bmatrix}^T + \\ & \frac{1}{2} \begin{bmatrix} \overline{k}_1^+ & 0 & 0 \\ 0 & \overline{k}_2^+ & 0 \\ 0 & 0 & \overline{k}_3^+ \end{bmatrix} - \frac{1}{2} \begin{bmatrix} \overline{k}_1^+ \\ \overline{k}_2^+ \\ \overline{k}_3^+ \end{bmatrix} \overline{N} \begin{bmatrix} \overline{k}_1^+ \\ \overline{k}_2^+ \\ \overline{k}_3^+ \end{bmatrix}^T + \\ & \frac{1}{2} \begin{bmatrix} -\overline{k}_1^- & 0 & 0 \\ 0 & -\overline{k}_2^- & 0 \\ 0 & 0 & -\overline{k}_3^- \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -\overline{k}_1^- \\ -\overline{k}_2^- \\ -\overline{k}_3^- \end{bmatrix} \overline{N} \begin{bmatrix} -\overline{k}_1^- \\ -\overline{k}_2^- \\ -\overline{k}_3^- \end{bmatrix}^T \end{aligned}$$

From the energy analysis, we know that $\overline{D}^{N-\mathcal{QRD}}$ defines a dissipative operator. However, in the nonlinear case, the additional diagonal matrix defining $M^{N-\mathcal{QRD}}$ does not cancel when assembling the global entropy balance, due to the variation of \overline{a} in space. Hence, the local entropy production of the scheme is

$$\begin{aligned} \Phi_{\mathcal{H}}^{N-\mathcal{QRD}} &= \frac{1}{2} \sum_{j \in E} v_j \overline{k}_j v_j + \epsilon_{\mathcal{H}}^{N-\mathcal{QRD}}, \\ \epsilon_{\mathcal{H}}^{N-\mathcal{QRD}} &= \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}^T \overline{D}^{N-\mathcal{QRD}} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \geq 0 \end{aligned} \quad (6.24)$$

One can now prove that the scheme is entropy consistent. In view of an important remark related to the \mathcal{CRD} variant of the N scheme, we report here one key element of the proof, due to [4]. In particular, we recall the following lemma (see [4, 9]).

Lemma 6.3.6 (Abgrall and Barth, 2002). *Given a bounded time $t_f > 0$, with $t_f = N\Delta t$, and the bounded sequence of continuous piecewise linear functions $v_h(x, y, t)$ with $v_h(x, y, t^n) = v_h^n$ such that*

$$\sup_h \sup_{(x,y,t)} \|v_h(x, y, t)\| \leq C, \quad \lim_{h \rightarrow 0} \|v_h - v\|_{L^2_{loc}(\Omega \times [0, t_f])} = 0$$

with C a constant independent of h and Δt , then

$$\lim_{h \rightarrow 0} \sum_{n=0}^N \Delta t \sum_{E \in \mathcal{T}_h} |E| \sum_{i,j \in E} \|v_i^n - v_j^n\| = 0$$

In particular, given $\sigma(1, 2, 3)$, a circular permutation of the indices of the nodes of an element E , the two piecewise constant functions

$$v'_h = \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} \chi_{S_i \cap E} v_i^n$$

and

$$\tilde{v}_h = \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} \chi_{S_i \cap E} v_{\sigma(i)}^n$$

converge to the same limit v of v_h , as $h \rightarrow 0$.

Note that the quantity χ_S used in the definition of v'_h and \bar{v}_h is the characteristic function (3.4). Using the last result, one can prove that

Proposition 6.3.7 (\mathcal{QRD} N scheme and entropy dissipation). *The \mathcal{QRD} N scheme with exact integration of the flux Jacobian is entropy dissipative. In particular:*

$$\sum_{E \in \mathcal{T}_h} \frac{1}{2} \sum_{j \in E} v_j \bar{k}_j v_j \xrightarrow{h \rightarrow 0} \int_{\Omega} \nabla \cdot \mathcal{G}(v) \, dx \, dy$$

Proof. See [4, 9]. □

The extension of the result to the \mathcal{QRD} case with approximate mean-value Jacobian linearization is given in [4]. For the \mathcal{CRD} N scheme things are different, since (6.24) no longer holds. However, using (6.10) and (5.50), we can write

$$\Phi_{\mathcal{H}}^{\text{N-}\mathcal{CRD}} = \int_E \bar{v}_h \nabla \cdot \mathcal{F}(v_h) \, dx \, dy + \epsilon_{\mathcal{H}}^{\text{N-}\mathcal{CRD}}, \quad \epsilon_{\mathcal{H}}^{\text{N-}\mathcal{CRD}} \geq 0 \quad (6.25)$$

with

$$\epsilon_{\mathcal{H}}^{\text{N-}\mathcal{CRD}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}^T D^{\text{N-}\mathcal{CRD}} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \geq 0$$

and the symmetric dissipation matrix $D^{\text{N-}\mathcal{CRD}}$ given by (see (5.50))

$$D^{\text{N-}\mathcal{CRD}} = \begin{bmatrix} k_1^+ & 0 & 0 \\ 0 & k_2^+ & 0 \\ 0 & 0 & k_3^+ \end{bmatrix} - \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix} N \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix}^T$$

In equation (6.25), we have introduced the piecewise constant function

$$\bar{v}_h = \sum_{E \in \mathcal{T}_h} \chi_E \sum_{j \in E} \beta_j^{\text{LDA}} v_j$$

The reason for writing $\Phi_{\mathcal{H}}^{\text{N-}\mathcal{CRD}}$ as in (6.25) and for reporting lemma 6.3.6 is that, by using this lemma and condition (5.4), one could prove that

$$\bar{v}_h \xrightarrow{h \rightarrow 0} v$$

which would constitute a proof of the fact that the \mathcal{CRD} N scheme is entropy dissipative. The reason why we do not put this assertion in a formal proposition is that as we can use lemma 6.3.6 to show that $\bar{v}_h \rightarrow v$ for \bar{v}_h defined using the distribution coefficients of the LDA scheme, we could prove similar propositions for any scheme to which we add or subtract entropy dissipative terms corresponding to some \mathcal{LP} scheme. Even though the arguments to show that the \mathcal{CRD} N scheme is dissipative seem correct, some details are missing or not well understood. Nevertheless, (6.25) shows that a degree of entropy dissipation is indeed present in the discretization.

6.3.3 Time integration

We make in this section a short digression related to the entropy dissipative properties of the fully discrete equations obtained when integrating (6.16) in time. In order to do this, we note that given two discrete time levels t^n and t^{n+1} , one can write [171],

$$\mathcal{H}_i^{n+1} - \mathcal{H}_i^n = \int_{t^n}^{t_{n+1}} \frac{d\mathcal{H}(v_i(\tau))}{d\tau} d\tau$$

Using the definition of v , and making the change of variable

$$\xi = \frac{t - t^{n+1/2}}{\Delta t}, \quad t^{n+1/2} = \frac{t^n + t^{n+1}}{2}$$

we have the two equivalent expressions

$$\mathcal{H}_i^{n+1} - \mathcal{H}_i^n = \int_{-1/2}^{+1/2} v_i(\xi) \frac{du(v_i(\xi))}{d\xi} d\xi = \int_{-1/2}^{+1/2} v_i(\xi) a_0(v_i(\xi)) \frac{dv_i(\xi)}{d\xi} d\xi \quad (6.26)$$

Fully discrete entropy balances can be derived by a proper choice of $v_i(\xi)$ and by simple manipulations of (6.26).

6.3.3.1 Explicit FE time-integration

Taking in (6.26)

$$v_i(\xi) = \frac{v_i^{n+1} + v_i^n}{2} + \xi(v_i^{n+1} - v_i^n)$$

one easily shows that

$$v_i^n(u_i^{n+1} - u_i^n) = \mathcal{H}_i^{n+1} - \mathcal{H}_i^n - \epsilon_i^{\text{FE}}$$

with ϵ_i^{FE} the entropy production in time of the forward Euler scheme, given by

$$\epsilon_i^{\text{FE}} = \int_{-1/2}^{+1/2} \left(\frac{1}{2} + \xi \right) (v_i^{n+1} - v_i^n) a_0(v_i(\xi)) (v_i^{n+1} - v_i^n) d\xi \geq 0 \quad (6.27)$$

As a consequence, the fully discrete \mathcal{RD} entropy balance becomes:

$$\sum_{i \in \mathcal{T}_h} |S_i| (\mathcal{H}_i^{n+1} - \mathcal{H}_i^n) = -\Delta t \sum_{E \in \mathcal{T}_h} \Phi_{\mathcal{H}}^E(v_h^n) + \sum_{i \in \mathcal{T}_h} |S_i| \epsilon_i^{\text{FE}} \quad (6.28)$$

As in the linear case (see prop 4.2.8), the explicit time integration adds an anti-dissipative terms to the entropy balance. As remarked in §4.2.1, the competing effects of the entropy dissipation of the spatial discretization and of the entropy production of the explicit scheme can be controlled by the magnitude of the time-step Δt . For entropy dissipative discretizations, one may then look for a limiting value of the time-step guaranteeing the dissipative character of the fully discrete equations. A discussion on the topic can be found in [171].

6.3.3.2 Implicit BE time-integration

With the same choice of $v_i(\xi)$ done for the FE scheme, one can show that

$$v_i^{n+1}(u_i^{n+1} - u_i^n) = \mathcal{H}_i^{n+1} - \mathcal{H}_i^n + \epsilon_i^{\text{BE}}$$

with ϵ_i^{BE} the entropy dissipation in time of the backward Euler scheme, given by

$$\epsilon_i^{\text{BE}} = \int_{-1/2}^{+1/2} \left(\frac{1}{2} - \xi \right) (v_i^{n+1} - v_i^n) a_0(v_i(\xi)) (v_i^{n+1} - v_i^n) d\xi \geq 0 \quad (6.29)$$

As a consequence, the fully discrete \mathcal{RD} entropy balance becomes:

$$\sum_{i \in \mathcal{T}_h} |S_i| (\mathcal{H}_i^{n+1} - \mathcal{H}_i^n) = -\Delta t \sum_{E \in \mathcal{T}_h} \Phi_{\mathcal{H}}^E(v_h^{n+1}) - \sum_{i \in \mathcal{T}_h} |S_i| \epsilon_i^{\text{BE}} \quad (6.30)$$

As in the linear case (see prop 4.2.8), the fully implicit time integration adds an entropy removing term which contributes to stabilize the discretization.

6.3.3.3 Trapezium time scheme and \mathcal{CN} scheme

Note that combining (6.30) and (6.28) we easily obtain

$$\sum_{i \in \mathcal{T}_h} |S_i| (\mathcal{H}_i^{n+1} - \mathcal{H}_i^n) = -\frac{\Delta t}{2} \sum_{E \in \mathcal{T}_h} \left(\Phi_{\mathcal{H}}^E(v_h^n) + \Phi_{\mathcal{H}}^E(v_h^{n+1}) \right) + \sum_{i \in \mathcal{T}_h} |S_i| \epsilon_i^{\text{T}} \quad (6.31)$$

with ϵ_i^{T} the entropy production in time of the trapezium scheme, given by

$$\epsilon_i^{\text{T}} = \frac{1}{2} (\epsilon_i^{\text{FE}} - \epsilon_i^{\text{BE}}) = \int_{-1/2}^{+1/2} 2\xi (v_i^{n+1} - v_i^n) a_0(v_i(\xi)) (v_i^{n+1} - v_i^n) d\xi$$

Depending on the solution, the sign of ϵ_i^{T} can be either positive or negative, hence producing or dissipating entropy. A scheme with entropy conservation properties can instead be obtained by taking in (6.26)

$$u_i(\xi) = \frac{u_i^{n+1} + u_i^n}{2} + \xi(u_i^{n+1} - u_i^n), \quad v_i(\xi) = v(u_i(\xi))$$

leading to

$$\mathcal{H}_i^{n+1} - \mathcal{H}_i^n = \left(\int_{-1/2}^{1/2} v(u_i(\xi)) d\xi \right) (u_i^{n+1} - u_i^n) = v_i^{\mathcal{CN}} (u_i^{n+1} - u_i^n)$$

Evaluating the spatial residuals in $v_h^{\mathcal{CN}}$, we obtain a scheme which conserves entropy *in time*:

$$\sum_{i \in \mathcal{T}_h} |S_i| (\mathcal{H}_i^{n+1} - \mathcal{H}_i^n) = -\Delta t \sum_{E \in \mathcal{T}_h} \Phi_{\mathcal{H}}^E(v_h^{\mathcal{CN}}) \quad (6.32)$$

This Crank-Nicholson scheme is the only two-step time integration scheme which preserves the entropy dissipation characteristics of the spatial discretization.

6.4 Computational examples

We show here a few results obtained with the \mathcal{CRD} schemes introduced in this chapter. In particular, we consider again the solution of (6.1) with the exponential flux function $\mathcal{F} = (e^u, u)$. The definition of the problem and of the boundary conditions are the same as in §6.1. We solve the problem with \mathcal{CRD} N and LDA schemes and with the limited \mathcal{CRD} N scheme, obtained by applying (5.65) to the N scheme (6.10). The schemes are written in terms of the conserved variable u and the residual is computed using (6.6) with a second-order trapezium rule on each edge of the elements. The average Jacobian needed for the evaluation of the k_j parameters has been computed as in §6.1:

$$\bar{a}_E = \frac{1}{3} \sum_{j \in E} \bar{a}(u_j)$$

We integrate in time (5.5) using the explicit FE scheme with local time-stepping and (see equations (4.9) and (5.44))

$$\Delta t_i = 0.9 \frac{|S_i|}{\sum_{E \in \mathcal{D}_i} k_i^+} \quad \forall i \in \mathcal{T}_h$$

We compare the results obtained with the \mathcal{CRD} schemes on an irregular grid with $h = 0.015$, to the reference solution of §6.1, for completeness reported in figure 6.4.

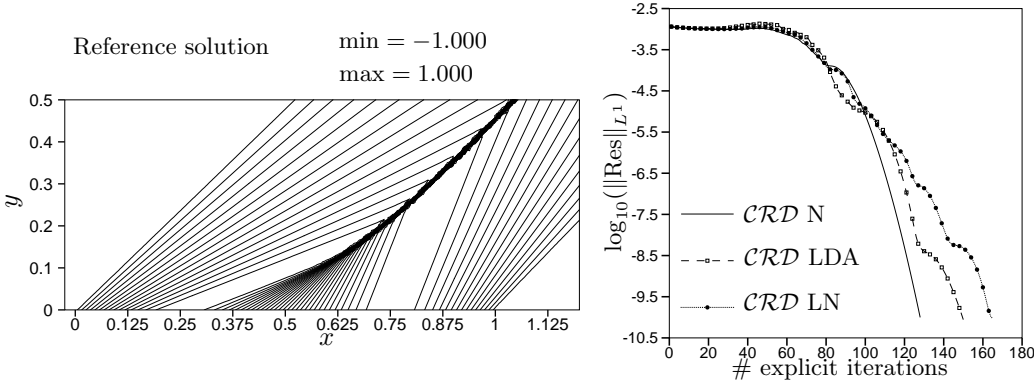


Figure 6.4: Nonlinear \mathcal{CL} with *exponential flux*. Reference solution (left) and convergence histories (right)

Recall that this solution has been computed on a finer unstructured mesh ($h = 1/200$) with the limited N scheme based on the \mathcal{QRD} formulation with 4 quadrature points in (6.4). As initial solution we set $u(x, y) = u(x, y = 0)$, with the latter as in (6.2). The convergence history of the \mathcal{CRD} schemes is reported on the right in figure 6.4, while the results are given in figures 6.5, 6.6 and 6.7. On the left, in the figures, we plot the contours of the steady solution u . On the right pictures, instead, we compare with the reference the distribution of u along the upper boundary. First of all, we remark that

no convergence problems are encountered and that all the schemes quickly reach the steady-state. By looking at the left picture on figure 6.5, we can see that the \mathcal{CRD} N scheme gives a solution free of numerical oscillations. In particular, the shock is monotonically captured, even though it is smeared over quite a few cells. The right picture further proves the non-oscillatory character of the computed shock and also it shows that its position and strength are correct, thus confirming the conservative character of the scheme.

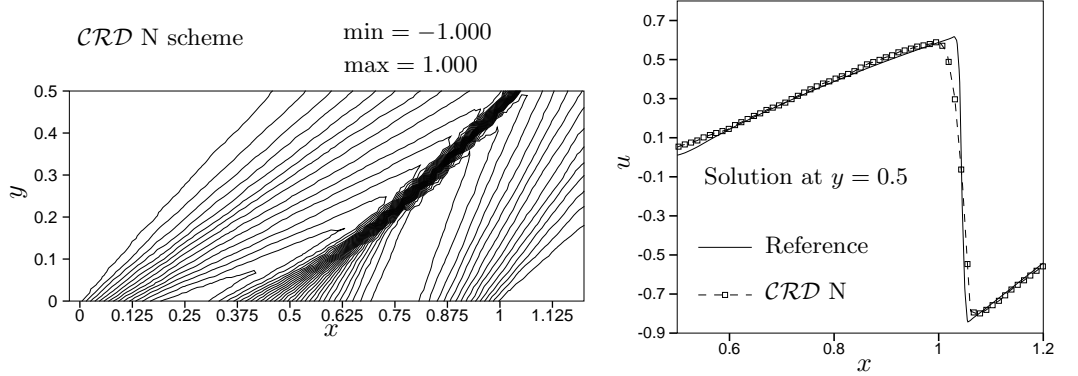


Figure 6.5: Nonlinear \mathcal{CL} with *exponential flux*. Solution obtained with the \mathcal{CRD} N scheme on the coarse mesh. Left: contour plot. Right: solution profile at $y = 0.5$

The numerical solution obtained with the \mathcal{CRD} LDA scheme is instead reported in figure 6.6. The left picture shows a less diffused numerical shock, however, as expected, oscillations are present in its proximity. The right picture shows the very sharp capturing of the discontinuity. Some small oscillations are also visible. Also for the LDA, position and strength of the shock are correct.

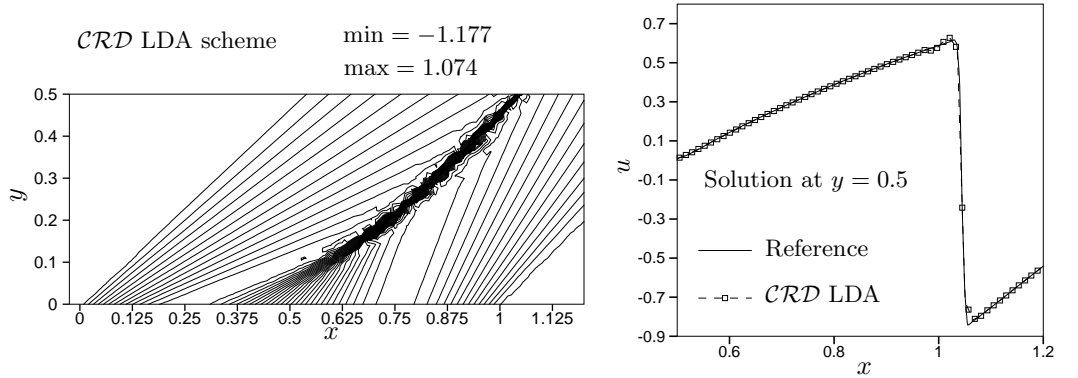


Figure 6.6: Nonlinear \mathcal{CL} with *exponential flux*. Solution obtained with the \mathcal{CRD} LDA scheme on the coarse mesh. Left: contour plot. Right: solution profile at $y = 0.5$

At last we comment the results obtained with the limited scheme, reported in figure 6.7. The contour plot shows a very crisp resolution of the shock. No oscillations are present in this result. The sharp approximation of the discontinuity is visible on the right picture. The resolution of the shock and of the smooth part of the solution are comparable with the ones obtained with the LDA scheme, except that the LN scheme yields an oscillation free result. The conservative character of the scheme is also confirmed.

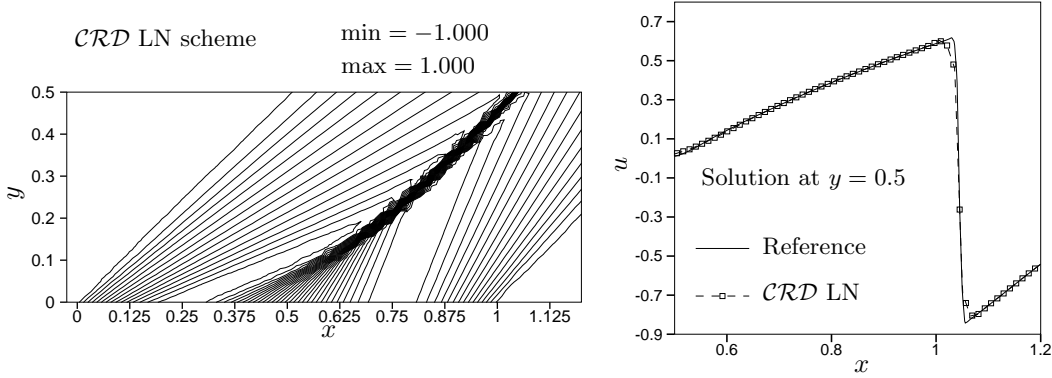


Figure 6.7: Nonlinear \mathcal{CL} with *exponential flux*. Solution obtained with the \mathcal{CRD} LN scheme on the coarse mesh. Left: contour plot. Right: solution profile at $y = 0.5$

6.5 Summary

This chapter has added another block to the construction of conservative schemes for nonlinear conservation laws. Its main contributions can be summarized as follows.

- With the help of a simple motivational example, we have introduced the need of a numerical discretization able to reproduce the correct jump conditions in correspondence of discontinuities: conservative \mathcal{RD} schemes;
- We have given a precise definition of a conservative \mathcal{RD} scheme and shown how this definition leads to an incompatibility between the need of the use of the \mathcal{CL} form of the problem, required to satisfy conservation, and the need of using the quasi-linear form of the equation to distribute the residual;
- We recalled the solution given in [4] to this incompatibility which has led to the definition of the \mathcal{QRD} approach. The main advantage of this approach is its formal equivalence with the linear case which allows to apply immediately most of the theory already developed;
- We have presented a simpler solution, namely the \mathcal{CRD} approach of [50]. This technique completely decouples the issues of conservation and upwinding by di-

rectly using the \mathcal{CL} form of the problem to define the local residual, while allowing to use an arbitrary local linearization of the quasi-linear form by forcing the \mathcal{RD} consistency condition of the schemes;

- The extension of basic \mathcal{RD} schemes to the \mathcal{CRD} framework has been presented. The key point of this framework, namely the \mathcal{CRD} formulation of the N scheme, has been discussed and compared to and other approaches. In particular, we have shown that the \mathcal{CRD} formulation becomes quite natural when writing the N scheme as the LDA plus dissipation. We also recalled the analogy with the conservative correction technique of [92, 48];
- We introduced an equivalence between the \mathcal{QRD} and the \mathcal{CRD} techniques. This *bridge* has allowed to show that, in general, the \mathcal{CRD} N scheme *does not* respect the sub-element LED condition;
- We have discussed the entropy stability of the schemes. A general semi-discrete \mathcal{RD} entropy balance has been introduced and used to define stable schemes;
- The analysis of the centered scheme has led to weaker definitions of stability: entropy consistency and entropy dissipation. Differently from entropy stability, entropy dissipation is related to the capability of a scheme of adding dissipation to a discrete approximation which converges to the integral of the divergence of the entropy flux when the mesh is refined;
- The entropy stability of \mathcal{MU} scheme has been studied. These schemes are entropy dissipative in the 1-target case. This shows the stabilizing effect of the \mathcal{MU} ;
- The cases of the LDA and N schemes have been considered in some detail. The analysis of the nonlinear case is quite similar to the energy stability analysis for the \mathcal{QRD} schemes, particularly for the LDA scheme. The \mathcal{CRD} formulation introduces complications due to the presence of the inexact flux Jacobian linearization. In the case of the N scheme, we have recalled the proof of its entropy dissipative character when an exact mean-value linearization of the Jacobians is used, due to [4]. Some elements that could allow the extension of the proof to the \mathcal{CRD} scheme have been given. However, details are still missing or not understood;
- The entropy dissipation introduced by the time integration scheme has been briefly analyzed after [171]. As in the linear case, the implicit BE scheme is the most stable, while the explicit FE scheme adds destabilizing terms to the discrete entropy balance. The entropy conservative \mathcal{CN} scheme has been presented;
- Illustrative computational results, involving the solution of a nonlinear \mathcal{CL} with an exponential flux, have been given. The results confirm the conservative character of the \mathcal{CRD} schemes. The non-oscillatory character of the \mathcal{CRD} N scheme and of its limited variant are also confirmed.

Chapter 7

Time dependent problems: conservative space-time \mathcal{RD}

This chapter will add the missing brick needed to approximate nonlinear conservation laws: a second-order non-oscillatory nonlinear scheme for time-dependent computations on unstructured grids. Starting from an improved prototype compact discretization for scalar advection we will arrive to a conservative formulation which, on one hand allows to solve general \mathcal{CL} s on unstructured meshes without the need of the introduction of complex mean-value linearizations of the flux Jacobians and, on the other hand, permits to construct well defined high-order nonlinear schemes which enjoy a true residual property and yield approximations of weak solutions free of numerical oscillations. Part of the material contained in this chapter is covered also in [7, 118, 8, 120, 47, 53, 44, 51] as far as scalar advection is concerned, and in [141, 142] concerning the extension to nonlinear conservation laws.

7.1 Time-dependent advection

We consider now the approximation of solutions of

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \mathcal{S}(x, y) \quad \text{on } \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (7.1)$$

in the time-dependent regime. Solutions of (7.1) are to be approximated on discretizations of Ω_T which, as described in §3.1, can be decomposed into space-time slabs $\Omega \times [t^n, t^{n+1}]$ in which Ω is discretized by means of an unstructured grid composed of triangular elements. The grid is denoted by \mathcal{T}_h . As in chapter 5, we will assume \vec{a} to be constant or, if $\vec{a} = \vec{a}(x, y)$, to be locally replaced by a proper average on each element $E \in \mathcal{T}_h$.

In §4.4.3, we have shown that if (7.1) is approximated by a scheme of the form (4.1), the spatial accuracy of the discretization is limited to first-order in time-dependent calculations. The objective of this section is then to present a more general compact prototype, which overcomes this limitation. All the schemes discussed in the previous chapter are particular cases of this abstract discrete model but, as remarked, are only first-order accurate in space. Second-order schemes encompassed by this new prototype are only of the \mathcal{FE} and \mathcal{FS} type. Second-order accurate \mathcal{FV} schemes, as well as higher order ones, do not fit in this framework due to the non-compact character of the reconstructions used to increase the accuracy. An exception to this is perhaps given by the Spectral Volume schemes developed in [182, 183, 184, 185], which however we do not consider here. One form of the \mathcal{RD} schemes which fits into the new framework will then be presented: the space-time schemes. The extension of the \mathcal{MU} schemes of chapter 5 is presented and discussed. Most of the presentation is done in the homogeneous case $\mathcal{S} = 0$. However, the extension to the general case is also underlined.

7.1.1 An improved prototype for unsteady simulations

In the homogeneous case, we will consider here schemes approximating (7.1) which can be recast in the following semi-discrete form:

$$\sum_{E \in \mathcal{D}_i} \sum_{j \in E} m_{ij}^E \frac{du_j}{dt} = - \sum_{E \in \mathcal{D}_i} \phi_i^E = - \sum_{E \in \mathcal{D}_i} c_{ij}^E (u_i - u_j) \quad (7.2)$$

Compared to (4.1), this prototype adds a coupling *in space* of the time derivatives of the nodal values, though the matrix elements m_{ij}^E . In the following we will refer to this matrix as to the *mass-matrix*, and denote it by $M^{d\tau}$. As done in chapter 4, we assume that schemes of the form (7.2) verify the following local consistency condition.

Assumption (Local Consistency - time-dependent case). *For a given scheme of the form (7.2), it is possible to find a consistent approximation of the unknown $u_h(x, y)$, and of the flux $\mathcal{F}_h(x, y) = (\vec{a}u)_h(x, y)$, such that $\forall E \in \mathcal{T}_h$*

$$\sum_{i \in E} \left(\sum_{j \in E} m_{ij}^E \frac{du_j}{dt} + \phi_i^E \right) = \int_E \left(\frac{\partial u_h}{\partial t} + \nabla \cdot \mathcal{F}_h \right) dx dy \quad (7.3)$$

One can easily show that this consistency assumption leads to

$$\sum_{i \in E} \beta_i^{d\tau} = 1, \quad \beta_i^{d\tau} = \frac{1}{|E|} \sum_{j \in E} m_{ij}^E \quad (7.4)$$

As usual, the properties of scheme (7.2) are characterized by the L^∞ and energy stability of the discrete solutions, as well as by the accuracy of the approximation. Concerning the L^∞ stability, one can argue that if the right hand side of (7.2) defines a LED scheme (in the sense of proposition 4.1.1), and if the mass-matrix is an \mathcal{M} -matrix, then, upon its inversion, one would end up with a scheme which is still LED, due to

the positivity of the entries of the inverse of $M^{d\tau}$. This can be used to characterize the energy stability of the discretization, as well. Of course, as a particular case we can consider the LED schemes presented in the previous chapters, for which the mass-matrix is diagonal with positive entries. However, a general characterization of the stability of the solutions of (7.2) is harder than for (4.1) and can be done on a case by case basis.

While no general results are reported concerning the stability of (7.2), we can characterize its accuracy in a general fashion. In particular, proceeding as in §4.4.1, §4.4.2 and §4.4.3, one can prove that

Proposition 7.1.1. *A scheme of the form (7.2) verifying the local consistency (7.3) for a continuous second-order accurate approximation of the flux \mathcal{F}_h and of the unknown u_h , is second-order accurate in space if*

$$\sum_{j \in E} m_{ij}^E \frac{du_j}{dt} + \phi_i^E = \mathcal{O}(h^3) \quad (7.5)$$

Proof. Omitted (see §4.4.1, §4.4.2 and §4.4.3). \square

It is clear that the extra coupling introduced by the mass-matrix allows to construct schemes satisfying (7.5). We give hereafter two examples of such schemes. The first shows that (4.1) encompasses \mathcal{FE} discretizations of (7.1). We then present a whole class of \mathcal{RD} schemes satisfying the accuracy requirement. Starting from the last example we will then introduce the space-time framework used in the thesis. We recall that, while the introduction of the mass-matrix allows to overcome the accuracy limitation, it does not allow to overcome the limitations imposed by Godunov's theorem 4.4.5. In particular, also for (7.2), *second-order schemes which also have some form of L^∞ stability cannot be linear*. The construction of nonlinear schemes, however, can be done only after introducing also the time-discretization. In our case, this is achieved in the above mentioned space-time \mathcal{RD} framework.

7.1.1.1 Finite element schemes with mass-matrix

A well-known member of the family of schemes defined by (7.2) is the Galerkin \mathcal{FE} scheme obtained as (see §5.3.1, equation (5.23)):

$$\int_{\Omega} \psi_i \frac{\partial u_h}{\partial t} dx dy + \int_{\Omega} \psi_i \vec{a} \cdot \nabla u_h dx dy = 0, \quad \forall i \in \mathcal{T}_h \quad (7.6)$$

with u_h as in (3.7). After using the properties of the basis functions ψ_i (equation (3.6)), recalling the analogy of §5.3.1 and using (5.7), we end up with a scheme formally identical to (4.1) with

$$c_{ij}^E = c_{ij}^G = -\frac{k_j}{3}, \quad m_{ij}^E = m_{ij}^G = \frac{|E|}{12} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

One easily checks that the scheme satisfies (7.3) and (7.4). The Galerkin scheme can be shown to respect (7.5) and has no L^∞ stability properties, as in the steady case. As before, its energy analysis is quite natural, since by construction one has

$$0 = \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} \sum_{j \in E} u_i m_{ij}^G \frac{du_j}{dt} + \sum_{i \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_i} u_i \phi_i^G = \int_{\Omega} u_h \frac{du_h}{dt} dx dy + \int_{\Omega} u_h \vec{a} \cdot \nabla u_h dx dy$$

which finally leads to

$$\int_{\Omega} \frac{d\mathcal{E}_h}{dt} dx dy + \frac{1}{2} \oint_{\partial\Omega} \mathcal{E}_h \vec{a} \cdot \hat{n} dl = 0$$

showing that the Galerkin scheme does not dissipate energy. The SUPG \mathcal{FE} scheme can be derived in a similar fashion. In particular, $\forall i \in \mathcal{T}_h$ one has

$$\int_{\Omega} \psi_i \frac{\partial u_h}{\partial t} dx dy + \int_{\Omega} \psi_i \vec{a} \cdot \nabla u_h dx dy + \sum_{E \in \mathcal{T}_h} \int_E \tau \vec{a} \cdot \nabla \psi_i \left(\frac{\partial u_h}{\partial t} + \vec{a} \cdot \nabla u_h \right) = 0 \quad (7.7)$$

As before, we obtain a scheme formally identical to (7.2) with

$$c_{ij}^{\text{SUPG}} = -\left(\frac{1}{3} + \tau \frac{k_i}{2|E|}\right) k_j, \quad m_{ij}^{\text{SUPG}} = \frac{1}{12} \begin{bmatrix} 2|E| + 2k_1 & |E| + 2k_1 & |E| + 2k_1 \\ |E| + 2k_2 & 2|E| + 2k_2 & |E| + 2k_2 \\ |E| + 2k_3 & |E| + 2k_3 & 2|E| + 2k_3 \end{bmatrix}$$

The satisfaction of (7.3) and (7.4) is easily verified. As the Galerkin scheme, the SUPG is second-order accurate and has no LED properties. Truly energy stable formulations are obtained more generally using the Least-Squares \mathcal{FE} approach for which we refer to [19] and references therein. We observe that both for the Galerkin scheme and for the SUPG scheme, formulation (4.1) is obtained by substituting to the mass-matrix of the schemes, the *lumped* mass-matrix obtained as

$$m_{ij}^{\text{lumped}} = \delta_{ij} \sum_{k \in E} m_{ik}^E = \delta_{ij} \frac{|E|}{3}$$

This mass lumping procedure clearly introduces an inconsistency, ultimately spoiling the spatial accuracy of the schemes, as confirmed by the analysis of §4.4.3.

7.1.1.2 A \mathcal{RD} Taylor-Galerkin approach: consistent LW scheme

In this section we show the construction of a consistent second-order cell-vertex Lax-Wendroff (LW) scheme on unstructured meshes. We will show that, if the hypothesis of linear variation of the solution is consistently taken into account, the scheme has a non-diagonal mass-matrix, thus giving further evidence of the claim made in §4.4.3, that the LW schemes proposed in [90, 60] are first-order accurate on general triangulations.

Following [129, 148], we perform the following Taylor expansion in time:

$$u^{n+1} = u^n + \left(\frac{\partial u}{\partial t} \right)^n \Delta t + \frac{\Delta t^2}{2} \left(\frac{\partial^2 u}{\partial t^2} \right)^n + \mathcal{O}(\Delta t^3)$$

Next, for linear scalar advection, one easily finds that

$$\frac{\partial u}{\partial t} = -\nabla \cdot (\vec{a} u) \quad \text{and} \quad \frac{\partial^2 u}{\partial t^2} = \nabla \cdot (\vec{a} \nabla \cdot (\vec{a} u))$$

hence

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

Last expression is a semi-discrete second-order accurate equivalent of the advection equation. Neglecting the terms of $\mathcal{O}(\Delta t^2)$, and discretizing the resulting expression with a Galerkin \mathcal{FE} approach leads to the well-known Taylor-Galerkin scheme [66].

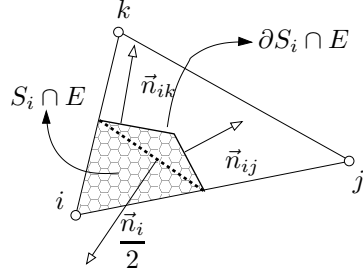


Figure 7.1: LW scheme: geometry of the construction

The \mathcal{FS} analog of the Taylor-Galerkin scheme in a generic node i , is instead obtained upon integration of the last expression over S_i :

$$\int_{S_i} \frac{u^{n+1} - u^n}{\Delta t} dx dy + \int_{S_i} \nabla \cdot (\vec{a} u)^n dx dy - \frac{\Delta t}{2} \int_{S_i} \nabla \cdot (\vec{a} \nabla \cdot (\vec{a} u))^n dx dy = 0$$

We then rewrite the last term on the left hand side (LHS) as a line integral, and express all the terms as the sum of contributions coming from the elements $E \in \mathcal{D}_i$:

$$\sum_{E \in \mathcal{D}_i} \left(\int_{S_i \cap E} \frac{u^{n+1} - u^n}{\Delta t} dx dy + \int_{S_i \cap E} \nabla \cdot (\vec{a} u)^n dx dy - \frac{\Delta t}{2} \oint_{\partial S_i \cap E} \nabla \cdot (\vec{a} u)^n \vec{a} \cdot \hat{n} dl \right) = 0$$

The second integral on the LHS is easily evaluated. If u^n is the second-order continuous piecewise linear approximation given by (3.7), one has, due to the definition of S_i

$$\nabla \cdot (\vec{a} u)^n|_E = \frac{\phi^h}{|E|} \implies \int_{S_i \cap E} \nabla \cdot (\vec{a} u)^n dx dy = \frac{1}{3} \phi^h(u^n)$$

with $\phi^h(u^n)$ as in (5.7). With reference to figure 7.1, the last term on the LHS becomes

$$\frac{\Delta t}{2} \oint_{\partial S_i \cap E} \nabla \cdot (\vec{a} u)^n \vec{a} \cdot \hat{n} dl = \frac{\Delta t}{2} \frac{\phi^h}{|E|} \vec{a} \cdot (\vec{n}_{ik} + \vec{n}_{ij}) = -\frac{\Delta t}{2|E|} \phi^h \frac{\vec{a} \cdot \vec{n}_i}{2} = -\frac{\Delta t}{2|E|} k_i \phi^h(u^n)$$

Finally, we have arrived to the LW scheme [129, 148]

$$\int_{S_i} \frac{u^{n+1} - u^n}{\Delta t} dx dy + \sum_{E \in \mathcal{D}_i} \beta_i^{\text{LW}} \phi^h(u^n) = 0, \quad \beta_i^{\text{LW}} = \frac{1}{3} + \frac{\Delta t}{2|E|} k_i \quad (7.8)$$

However, we still have to evaluate the first integral. Using the consistent linear variation of u^{n+1} and u^n in space, given by (3.7), this term becomes

$$\int_{S_i} \frac{u^{n+1} - u^n}{\Delta t} dx dy = \sum_{E \in \mathcal{D}_i} \int_{S_i \cap E} \frac{u^{n+1} - u^n}{\Delta t} dx dy = \sum_{E \in \mathcal{D}_i} \sum_{j \in E} m_{ij}^{\text{LW}} \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

where the *consistent \mathcal{RD} Lax-Wendroff mass-matrix* m_{ij}^{LW} is given by

$$m_{ij}^{\text{LW}} = \frac{|E|}{108} \begin{bmatrix} 22 & 7 & 7 \\ 7 & 22 & 7 \\ 7 & 7 & 22 \end{bmatrix} \quad (7.9)$$

Finally, the second-order accurate \mathcal{RD} LW scheme reads

$$\sum_{E \in \mathcal{D}_i} \left(m_{ij}^{\text{LW}} \frac{u_j^{n+1} - u_j^n}{\Delta t} + \beta_i^{\text{LW}} \phi^h(u^n) \right) = 0, \quad \forall i \in \mathcal{T}_h \quad (7.10)$$

with m_{ij}^{LW} , β_i^{LW} and $\phi^h(u^n)$ as in (7.9), (7.8), and (5.7), respectively. The LW scheme of [90, 60] is obtained from this consistent discretization upon the lumping of m_{ij}^{LW} , yielding

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|S_i|} \sum_{E \in \mathcal{D}_i} \beta_i^{\text{LW}} \phi^h(u^n), \quad \forall i \in \mathcal{T}_h \quad (7.11)$$

As in the case of the Galerkin and SUPG schemes, the lumping on the mass-matrix generally leads to an inconsistent first-order discretization. The success of this inconsistent LW scheme is attributed to an error cancellation taking place on structured meshes, on which most accuracy studies have been presented in literature. For the structured triangulation on the right on figure 3.1, this error cancellation has been shown in [143], where, using Taylor series expansions in space, it has been proved that the inconsistent LW scheme respects the modified equation

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \frac{\|\vec{a}\| h^2}{6} \left(\nu^2 \hat{a} \cdot (\hat{a} \cdot (\nabla \cdot (\hat{a} \cdot \nabla u))) - \hat{a} \cdot \nabla \left(\frac{\partial^2 u}{\partial x \partial y} + \Delta u \right) \right) + \mathcal{O}(h^3, \Delta t^3)$$

with $\Delta(\cdot)$ the Laplacian operator, $\hat{a} = \vec{a}/\|\vec{a}\|$ and with ν the CFL parameter

$$\nu = \frac{\|\vec{a}\| \Delta t}{h}$$

This modified equation shows the second-order of accuracy of the inconsistent scheme on the mesh of figure 3.1, and in one space dimension reduces to the well known LW modified equation [110]

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = -\frac{|a| h^2}{6} (1 - \nu^2) \frac{\partial^3 u}{\partial x^3} + \mathcal{O}(h^3, \Delta t^3)$$

However, on general meshes, the error cancellation leading to this result will not occur. This analysis show that on general meshes the inconsistent LW scheme (7.11) is not a second-order discretization. We have to mention that this is in contraddiction with some results published very recently in [62]. The reference contains a grid convergence study on irregular meshes showing second-order of accuracy for scheme (7.11). The construction of this section fails to justify the result of the reference. We also recall that, in a different framework, the mass matrix m_{ij}^{LW} was already presented in [38].

7.1.1.3 \mathcal{FS} schemes for time-dependent computations

Early attempts to extend the use of \mathcal{RD} schemes to the simulation of time-dependent problems have resorted to the analogy with Petrov-Galerkin \mathcal{FE} schemes. This has been done for example in [113, 71] and later in [7, 8, 120]. More recently, a general family of mass-matrices for \mathcal{RD} schemes, which encompasses the PG mass-matrix, has been introduced in [62, 61]. Here, we discuss an approach that we believe being the most faithful to the \mathcal{FS} spirit. The technique we are going to introduce is at the basis of the work of D. Caraei and collaborators [33, 32, 34, 35, 36, 37, 38] and leads to the class of second-order schemes that will be considered later.

The idea at the basis of this approach is that, given the second-order accurate approximations u_h and \mathcal{F}_h , one easily shows that for a smooth solution u

$$\int_E \left(\frac{\partial u_h}{\partial t} + \nabla \cdot \mathcal{F}_h \right) dx dy = \int_E \left(\frac{\partial(u_h - u)}{\partial t} + \nabla \cdot (\mathcal{F}_h - \mathcal{F}) \right) dx dy = \mathcal{O}(h^3)$$

As a consequence, second-order schemes can be obtained just by distributing with uniformly bounded distribution coefficients the residual obtained by integrating the whole equation, time derivative included:

Proposition 7.1.2 (\mathcal{LP} schemes - time-dependent case). *Given a second-order accurate approximation of the unknown u_h and of the flux \mathcal{F}_h , Linearity Preserving \mathcal{RD} schemes are second-order accurate in space in time-dependent computations, provided that the element residual is defined as*

$$\phi^h = \int_E \left(\frac{\partial u_h}{\partial t} + \nabla \cdot \mathcal{F}_h \right) dx dy$$

In particular, using (3.7) and the properties of the basis functions (3.6), one easily arrives to a scheme of the form (7.2) with

$$c_{ij}^{\mathcal{LP}} = -\beta_i k_j, \quad m_{ij}^{\mathcal{LP}} = \frac{E}{3} \begin{bmatrix} \beta_1 & \beta_1 & \beta_1 \\ \beta_2 & \beta_2 & \beta_2 \\ \beta_3 & \beta_3 & \beta_3 \end{bmatrix}$$

While this approach allows to *recover* the accuracy in space, the overall accuracy will depend on how the time derivative is discretized. However, now a truly second-order discretization can be obtained by selecting a time-integration scheme with accuracy at least equal to two. The problem is to define properly the distribution coefficients β_j , in order to be able to ensure, in unsteady computations, some form of L^∞ stability. To do this, two things are necessary: a time-integration scheme and a technique to construct (L^∞ -)stable nonlinear schemes. The first element, is needed because, if we are to make use of the theory of positive coefficients, we need to study the properties of the whole discretization, as the results of §4.1.1 and §4.1.2 show. The second element is of course related to the fact that a high-order scheme has to be nonlinear to be also stable, as stated by Godunov's theorem. The development of such a construction is the topic of the next sections.

7.1.2 A space-time framework

We construct in this section fully discrete analogs of (7.1). As always, this is done in a generic space-time slab $\Omega \times [t^n, t^{n+1}]$. In particular, we note that any element $E \in \mathcal{T}_h$, defines in $\Omega \times [t^n, t^{n+1}]$ a prismatic subset, as depicted on figure 7.2. Using the notation of §3.1 and §3.2, we are interested in discretizations approximating time-dependent solutions of (7.1) in $\Omega \times [t^n, t^{n+1}]$ as follows.

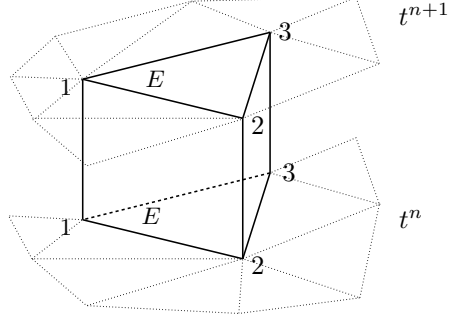


Figure 7.2: Space-time prism in $\Omega \times [t^n, t^{n+1}]$

Definition 7.1.3 (Space-time \mathcal{RD} scheme). A space-time Residual Distribution or Fluctuation Splitting scheme is defined as one that, given u^n the continuous approximation of u at time t^n (equations (3.7) and (3.9)), given the continuous space-time approximation of the unknown $u^h(x, y, t)$ (equation (3.9)) with $u^h(x, y, t^n) = u^n$, and given the continuous approximation of the flux \mathcal{F}^h (3.14), computes $u^h(x, y, t^{n+1}) = u^{n+1}$ as follows:

1. $\forall E \in \mathcal{T}_h$ compute on the space-time prism $E \times [t^n, t^{n+1}]$ the residual

$$\phi^{n+1} = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u^h}{\partial t} + \nabla \cdot \mathcal{F}^h \right) dx dy dt \quad (7.12)$$

2. $\forall E \in \mathcal{T}_h$ distribute fractions of ϕ^{n+1} to the nodes of E . Denoting by ϕ_i^{n+1} the split residual or local nodal residual for node $i \in E$, one must have by construction

$$\sum_{j \in E} \phi_j^{n+1} = \phi^{n+1} = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u^h}{\partial t} + \nabla \cdot \mathcal{F}^h \right) dx dy dt \quad (7.13)$$

Equivalently, denoting by β_i the distribution coefficient of node i :

$$\beta_i = \frac{\phi_i^{n+1}}{\phi^{n+1}} \quad (7.14)$$

one must have by construction

$$\sum_{j \in E} \beta_j = 1 \quad (7.15)$$

3. $\forall i \in \mathcal{T}_h$ assemble the elemental contributions of all $E \in \mathcal{D}_i$ and compute the nodal values of u^{n+1} by solving the algebraic system

$$\sum_{E \in \mathcal{D}_i} \phi_i^{n+1} = 0, \quad \forall i \in \mathcal{T}_h \quad (7.16)$$

As we will see, all the schemes of type (7.2), can be recast in this space-time \mathcal{RD} framework if the time derivative is integrated with the trapezium scheme. However, our first objective is to characterize the accuracy of the space-time schemes.

7.1.2.1 Accuracy of space-time \mathcal{RD}

The analysis of the accuracy of space-time residual distribution is formally identical to what has been done in §4.4.1, §4.4.2 and §4.4.3. Consider then the solution of (7.1), where, as explained in §3.1, the temporal domain is discretized by means of M time levels $\{t^1 = 0, t^2, \dots, t^n, t^{n+1}, \dots, t^M = t_f\}$. We denote by Δt a characteristic value of the time-step, for example

$$\Delta t = \min_n (t^{n+1} - t^n)$$

Given a function $\varphi \in C_0^1(\Omega_T)$, with $\Omega_T = \Omega \times [0, t_f]$, we then evaluate the following expression:

$$0 = \sum_{n=1}^{M-1} \sum_{i \in \mathcal{T}_h} \varphi_i^{n+1} \sum_{E \in \mathcal{D}_i} \phi_i^{n+1} = \sum_{n=1}^{M-1} \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \varphi_i^{n+1} \phi_i^{n+1}$$

where $\varphi_i^{n+1} = \varphi(x_i, y_i, t^{n+1})$. We then introduce the continuous piecewise linear in space and time approximation φ^h , obtained interpolating the values φ_i^n and φ_i^{n+1} , as in (3.9). Proceeding as in §4.4.1, §4.4.2 and §4.4.3, using the uniform boundedness of $\|\nabla \varphi\|$ and $|\partial \varphi / \partial t|$ and the consistency relation (7.13), one easily shows that, up to terms of $\mathcal{O}(h^2, \Delta t^2)$, last expression can be rewritten as

$$\int_{\Omega_T} \varphi^h \left(\frac{\partial u^h}{\partial t} + \nabla \cdot \mathcal{F}^h \right) dx dy dt + \sum_{n=1}^{M-1} \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i^{n+1} - \bar{\varphi}) \phi_i^{n+1} = 0 \quad (7.17)$$

where, $\bar{\varphi}$ is the average value of φ^h

$$\bar{\varphi} = \frac{1}{\Delta t |E|} \int_{t^n}^{t^{n+1}} \int_E \varphi^h dx dy dt$$

Given a smooth exact solution u , we can evaluate the accuracy of the approximation by noting that (7.17) is equivalent to

$$\int_{\Omega_T} \varphi^h \left(\frac{\partial(u^h - u)}{\partial t} + \nabla \cdot (\mathcal{F}^h - \mathcal{F}) \right) dx dy dt + \sum_{n=1}^{M-1} \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i^{n+1} - \bar{\varphi}) \phi_i^{n+1} = 0$$

We remark now that, since $\varphi \in C_0^1(\Omega_T)$, hence $\varphi = 0$ on $\partial\Omega_T$ and both $\|\nabla \varphi^h\|$ and $|\partial \varphi^h / \partial t|$ are bounded uniformly with respect to h and Δt , we have

$$\begin{aligned} \int_{\Omega_T} \varphi^h \left(\frac{\partial(u^h - u)}{\partial t} + \nabla \cdot (\mathcal{F}^h - \mathcal{F}) \right) dx dy dt = \\ - \int_{\Omega_T} \left((u^h - u) \frac{\partial \varphi^h}{\partial t} + \nabla \varphi^h \cdot (\mathcal{F}^h - \mathcal{F}) \right) dx dy dt = \mathcal{O}(h^2, \Delta t^2) \end{aligned}$$

which is within the approximation introduced in writing (7.17), and shows that our scheme will be second-order accurate in space and time if

$$\sum_{n=1}^{M-1} \sum_{E \in \mathcal{T}_h} \sum_{i \in E} (\varphi_i^{n+1} - \bar{\varphi}) \phi_i^{n+1} = \mathcal{O}(h^2, \Delta t^2)$$

Since, the number of time levels is of $\mathcal{O}(\Delta t^{-1})$, while the total number of elements is of $\mathcal{O}(h^{-2})$, last condition reduces to

$$(\varphi_i^{n+1} - \bar{\varphi}) \phi_i^{n+1} = \mathcal{O}(h^4, \Delta t^3)$$

The regularity of φ implies that $\varphi_i^{n+1} - \bar{\varphi} = \mathcal{O}(h, \Delta t)$, so that we end up with the following accuracy condition.

Proposition 7.1.4 (Accuracy of space-time \mathcal{RD} schemes). *A space-time \mathcal{RD} scheme is second-order accurate in space and time if*

$$\phi_i^{n+1} = \mathcal{O}(h^3, \Delta t^2) \tag{7.18}$$

7.1.2.2 The space-time residual

The analysis of the space-time residual performed here allows to fulfill three main objectives. First, characterize a particular class of schemes respecting (7.18) by construction. Then, show that schemes of the form (7.2) fit into the space-time framework when trapezium time integration is used. Lastly, present an entire class of linear positive space-time schemes.

First we note that, for an exact smooth solution u , one has

$$\begin{aligned}\phi^{n+1} &= \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u^h}{\partial t} + \nabla \cdot \mathcal{F}^h \right) dx dy dt = \\ &= \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial(u^h - u)}{\partial t} + \nabla \cdot (\mathcal{F}^h - \mathcal{F}) \right) dx dy dt = \\ &= \mathcal{O}(h^4, \Delta t^2) + \mathcal{O}(h^3, \Delta t^3) = \mathcal{O}(h^3, \Delta t^2)\end{aligned}$$

Hence, schemes respecting (7.18) can be obtained by simply distributing the residual with uniformly bounded distribution coefficients. Which leads to the conclusion that

Proposition 7.1.5 (Space-time \mathcal{LP} schemes). *Linearity preserving space-time schemes are second-order accurate.*

For scalar advection, the linearity of the problem allows to give a formal expression of the residual. In particular, with the notation of §5.1.1, using the form of u^h and $\mathcal{F}^h = \vec{a}u^h$ and the properties of the basis functions (3.6), one easily shows that [118, 8]

$$\phi^{n+1} = \sum_{j \in E} \frac{E}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{j \in E} (k_j u_j^n + k_j u_j^{n+1}) \quad (7.19)$$

On the other hand, integration of (7.2) with the trapezium scheme (equivalent in this linear case to \mathcal{CN}), gives

$$\sum_{E \in \mathcal{D}_i} \sum_{j \in E} m_{ij}^E (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{E \in \mathcal{D}_i} (c_{ij}^E (u_i^n - u_j^n) + c_{ij}^E (u_i^{n+1} - u_j^{n+1})) = 0$$

Using the consistency relations (7.3) and (7.4), one immediately shows that

$$\sum_{i \in E} \left(\sum_{j \in E} m_{ij}^E (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{E \in \mathcal{D}_i} (c_{ij}^E (u_i^n - u_j^n) + c_{ij}^E (u_i^{n+1} - u_j^{n+1})) \right) = \phi^{n+1},$$

so that any scheme of the form (7.2), with trapezium time integration, can be seen as a space-time \mathcal{RD} scheme. In the linear case, this carries on to the space-time framework the analogy with \mathcal{FE} schemes. It also shows that, as a particular case of (7.2), first-order \mathcal{FV} schemes on the median dual cell can be recast in this formalism. More generally, it shows the existence of an entire family of positive space-time schemes:

Proposition 7.1.6 (Linear positive space-time schemes). *A positive linear space-time \mathcal{RD} scheme is obtained from a linear LED \mathcal{RD} one, upon integration of (5.5) with the trapezium scheme. The positivity of the resulting discretization is constrained by the time-step restrictions of propositions 4.1.7 and 4.1.8.*

As a last remark, we note that in the inhomogeneous case the source term is easily included in to the discretization. In particular, using the linear approximation \mathcal{S}_h of

equation (3.15), the definition of the schemes remains unchanged except that now

$$\phi^{n+1} = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u^h}{\partial t} + \nabla \cdot \mathcal{F}^h - \mathcal{S}_h \right) dx dy dt$$

As before, one easily shows that in the linear case

$$\phi^{n+1} = \sum_{j \in E} \frac{|E|}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{j \in E} (k_j u_j^n + k_j u_j^{n+1}) - \Delta t \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (7.20)$$

7.1.3 Geometry of space-time \mathcal{RD} schemes

We derive an alternate expression of the residual, which will allow to give a geometrical interpretation of the meaning of ϕ^{n+1} and to define true space-time variants of the \mathcal{MU} schemes introduced in chapter 5. We start recasting (7.19) as

$$\phi^{n+1} = \sum_{j \in E} \left(\frac{\Delta t k_j}{2} + \frac{|E|}{3} \right) u_j^{n+1} + \sum_{j \in E} \left(\frac{\Delta t k_j}{2} - \frac{|E|}{3} \right) u_j^n$$

The weights multiplying the nodal values of u^h are the space-time \tilde{k}_j and \hat{k}_j parameters defined in §3.3 (equations (3.28) and (3.29)). Hence

$$\phi^{n+1} = \sum_{j \in E} \tilde{k}_j u_j^{n+1} + \sum_{j \in E} \hat{k}_j u_j^n$$

Introducing the *space-time flux* $(\vec{a}u, u) \in \mathbb{R}^2 \times \mathbb{R}$, we can show that the \tilde{k}_j and \hat{k}_j parameters are the projection of the *space-time flux Jacobian* $(\vec{a}, 1) \in \mathbb{R}^2 \times \mathbb{R}$ along directions determined by the geometry of the prism $E \times [t^n, t^{n+1}]$.

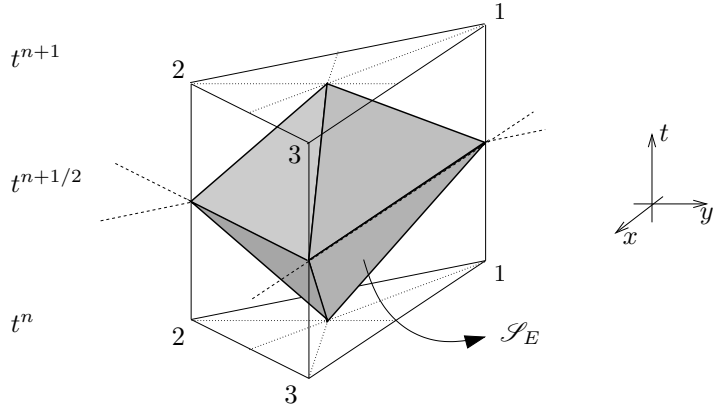


Figure 7.3: Closed shell in $E \times [t^n, t^{n+1}]$

To do this, we consider the shell \mathcal{S}_E formed by joining the gravity centers of E at times t^n and t^{n+1} with the nodes of the element at time $t^{n+1/2} = t^n + (t^{n+1} - t^n)/2$. As it can be seen from figure 7.3, this closed shell is all contained in the prism $E \times [t^n, t^{n+1}]$. We can associate to each node of the prism the face of \mathcal{S}_E opposite to it. This is illustrated on figure 7.4 for node 1. With reference to the figure, we introduce the space-time vectors \tilde{n}_1 and \hat{n}_1 , normal to the faces of \mathcal{S}_E opposite to node 1, pointing inward with respect to the shell, and scaled by the area of the faces.

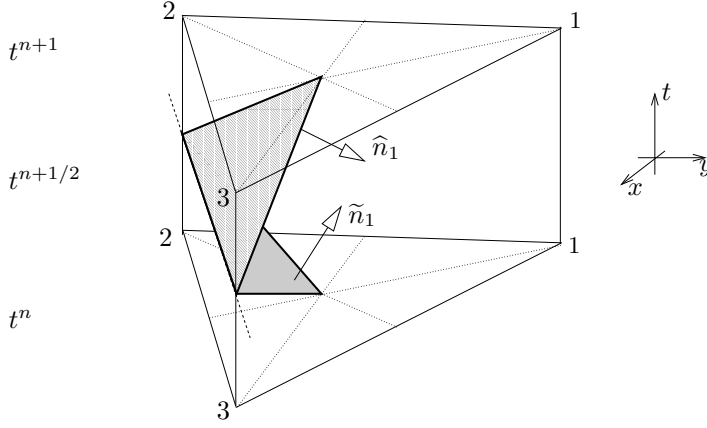


Figure 7.4: Space-time directions \tilde{n}_1 and \hat{n}_1 used to define \tilde{k}_1 and \hat{k}_1

Simple geometry shows that

$$\tilde{k}_1 = \tilde{n}_1 \cdot (\vec{a}, 1) \quad \text{and} \quad \hat{k}_1 = \hat{n}_1 \cdot (\vec{a}, 1)$$

Since $(\vec{a}, 1)$ determines the direction of a characteristic line cutting through the prism, we deduce that \tilde{k}_1 is the projection of the direction of the characteristic onto \tilde{n}_1 , and similarly for \hat{k}_1 . We know from the analysis of the advection equation (§2.1) that for the exact solution all the information propagates along $(\vec{a}, 1)$. We have now the possibility to apply this criterion to design schemes with a true space-time \mathcal{MU} character in which node 1 at time t^{n+1} receives a portion of ϕ^{n+1} only if $\tilde{k}_1 > 0$. This philosophy is at the basis of the space-time schemes proposed in [47, 53, 44, 51], the case of prismatic space-time elements being discussed in [51]. Clearly, the construction of figure 7.4 can be repeated for nodes 2 and 3. All the \tilde{k}_j and \hat{k}_j parameters can be written as the projection of the space-time flux Jacobian on directions which depend on the geometry of $E \times [t^n, t^{n+1}]$. If, as in the scalar case, we introduce *space-time inflow and outflow* states defined as

$$\tilde{u}_{in} = \sum_{j \in E} \left(\sum_{j \in E} (\tilde{k}_j^- + \hat{k}_j^-) \right)^{-1} (\tilde{k}_j^- u_j^{n+1} + \hat{k}_j^- u_j^n) = - \sum_{j \in E} \tilde{N} (\tilde{k}_j^- u_j^{n+1} + \hat{k}_j^- u_j^n), \quad (7.21)$$

and

$$\tilde{u}_{out} = \sum_{j \in E} \left(\sum_{j \in E} (\tilde{k}_j^+ + \hat{k}_j^+) \right)^{-1} (\tilde{k}_j^+ u_j^{n+1} + \hat{k}_j^+ u_j^n) = \sum_{j \in E} \tilde{N} (\tilde{k}_j^+ u_j^{n+1} + \hat{k}_j^+ u_j^n), \quad (7.22)$$

with

$$\tilde{N} = \left(\sum_{j \in E} (\tilde{k}_j^+ + \hat{k}_j^+) \right)^{-1}, \quad (7.23)$$

the residual can equivalently written as

$$\phi^{n+1} = \left(\sum_{j \in E} (\tilde{k}_j^+ + \hat{k}_j^+) \right) (\tilde{u}_{out} - \tilde{u}_{in}). \quad (7.24)$$

Last equations shows the analogy with a onedimensional balance along the characteristic line ζ intersecting the prism $E \times [t^n, t^{n+1}]$ in \tilde{u}_{out} and \tilde{u}_{in} . In particular, we remark that since the \hat{k}_j are not necessarily all negative, \tilde{u}_{in} does not necessarily lay on the plane $t = t^n$. Similarly, \tilde{u}_{out} does not necessarily lay on the plane $t = t^{n+1}$. In general, one will have a configuration as, for example, the one in figure 7.5. We will immediately see the implications of this in the design of space-time \mathcal{MU} schemes.

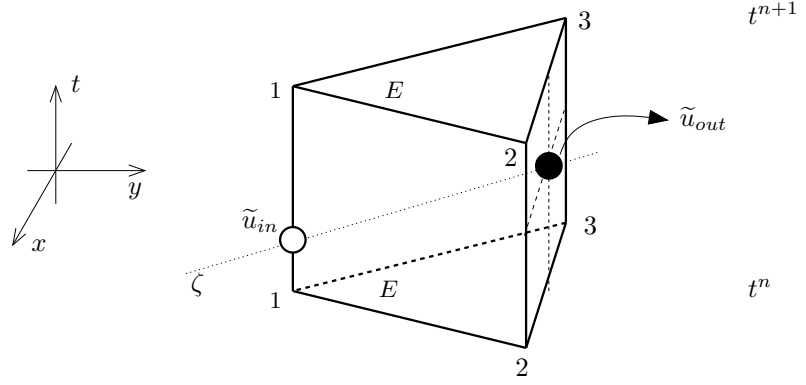


Figure 7.5: Space-time inflow and outflow states

7.1.4 Space-time \mathcal{MU} schemes

We introduce now the extension of some of the schemes presented in chapter 5, to the space-time framework. The analysis made in §7.1.2.2 gives already a means to obtain such an extension. Here, we will see more in detail how this works for some schemes. In particular, from now on, we will only consider the LDA, N and limited N schemes, since these are the ones we will use in the computations presented later. As seen in chapter 5, these schemes have, in the case of the steady advection equation, a \mathcal{MU} character. The analysis of the last section, however, shows that in the time-dependent case there is room for a different definition of upwinding, related to the space-time nature of the discretization.

In order to give this definition, we have to enlarge the class of schemes we consider. By construction, definition 7.1.3 gives a time-marching procedure allowing to compute the unknown at time t^{n+1} , given its nodal values at time t^n . Suppose instead to be

solving on the entire space-time domain at once, on a discretization which is given by the ensemble of the space-time prisms $E \times [t^n, t^{n+1}]$, $\forall E \in \mathcal{T}_h$ and $\forall n = 1, M$. In this case, the fully discrete analog of (7.1) becomes

$$\begin{aligned} \sum_{E \in \mathcal{D}_i} \phi_i^{n,n} + \sum_{E \in \mathcal{D}_i} \phi_i^{n,n+1} &= 0, & \forall i \in \mathcal{T}_h, \forall n = 2, M-1 \\ \sum_{E \in \mathcal{D}_i} \phi_i^{M,M} &= 0, & \forall i \in \mathcal{T}_h \end{aligned}$$

where $\forall E \in \mathcal{T}_h$ and $\forall n = 1, M-1$

$$\sum_{j \in E} (\phi_j^{n,n} + \phi_j^{n+1,n}) = \phi^n, \quad \sum_{j \in E} (\phi_j^{n,n+1} + \phi_j^{n+1,n+1}) = \phi^{n+1}$$

so that $\phi_i^{n,n+1}$ represents the fraction of ϕ^{n+1} distributed to node i at time t^n . A scheme will be space-time- \mathcal{MU} if

Definition 7.1.7 (Space-time- \mathcal{MU} scheme). A \mathcal{FS} scheme is space-time multidimensional upwind if in the prism $E \times [t^n, t^{n+1}]$

$$\begin{aligned} \tilde{k}_j^+ &= 0 \implies \phi_j^{n+1,n+1} = 0 \\ \hat{k}_j^+ &= 0 \implies \phi_j^{n,n+1} = 0 \end{aligned}$$

Proposition 7.1.8 (Space-time- \mathcal{MU} schemes and time-marching). A space-time- \mathcal{MU} scheme defines a time-marching procedure if

$$\Delta t = t^{n+1} - t^n \leq \min_{E \in \mathcal{T}_h} \min_{j \in E} \frac{2|E|}{3k_j^+}, \quad \forall n = 1, M-1 \quad (7.25)$$

Proof. Condition (7.25) ensures that $\hat{k}_j^+ = 0$ in all the elements of the mesh and in all space-time slabs. Hence, in every prism of every space-time slab $\Omega \times [t^n, t^{n+1}]$ a space-time- \mathcal{MU} scheme will never distribute any residual to the nodes at time t^n , hence decoupling the values of u^h in these nodes from its values at time t^{n+1} , thus yielding a true time-marching procedure. \square

In [47, 53, 44, 51], condition (7.25) is called the *past-shield* condition. Surprisingly enough, on prismatic space-time elements, the past-shield condition is exactly equivalent to the time-step restriction ensuring the local positivity of the N scheme with trapezium time integration¹ (see equation (5.45)). This condition allows to recast space-time- \mathcal{MU} schemes into the framework of definition 7.1.3. With the exception of §7.1.5, in the following we will always assume that (7.25) is satisfied. This also allows to simplify our notation. In particular, from now on we will denote by ϕ_i the portion of space-time residual distributed to node i in the space-time prism $E \times [t^n, t^{n+1}]$. No confusion is generated, since (7.25) guarantees that the characterization of definition 7.1.3 is valid, hence only the nodal values of u^{n+1} are to be computed in $\Omega \times [t^n, t^{n+1}]$. Also, in element $E \times [t^n, t^{n+1}]$ we will change the notation used for the residual to ϕ^h , so as to have uniform labeling with the previous chapters.

¹or equivalently \mathcal{CN} for a linear problem

7.1.4.1 LDA schemes

The \mathcal{LP} LDA scheme can be extended to the space-time framework in several different ways. As recalled in §7.1.1.2, early attempts to use the \mathcal{FS} approach for time-dependent computations resorted to the analogy with finite element PG schemes, thus introducing a consistent mass-matrix. For the LDA scheme this has been done in [113, 71] and later in [7, 8, 120]. Here, we will not consider this formulation. Instead, we will use the characterization of a space-time \mathcal{LP} scheme of proposition 7.1.5. In particular, not to introduce too much additional notation, in unsteady computations we simply refer to the LDA scheme as to the one defined by

$$\phi_i^{\text{LDA}} = k_i^+ N \phi^h = \beta_i^{\text{LDA}} \phi^h \quad (7.26)$$

with N as is (5.9) and ϕ^h given by (7.19). The LDA scheme is \mathcal{MU} but *not* space-time- \mathcal{MU} . In particular, as defined here it is equivalent to the scheme proposed in [33, 32, 34, 35, 36, 37, 38], except that a different time integration method is used in the references. We then consider the space-time- \mathcal{MU} analog of this scheme, the ST-LDA scheme defined by

$$\phi_i^{\text{ST-LDA}} = \tilde{k}_i^+ \tilde{N} \phi^h = \tilde{\beta}_i^{\text{ST-LDA}} \phi^h \quad (7.27)$$

where now, due to the satisfaction of (7.25), the parameter \tilde{N} is given by

$$\tilde{N} = \left(\sum_{j \in E} \tilde{k}_j^+ \right)^{-1} \quad (7.28)$$

Both the LDA and the ST-LDA schemes are second-order accurate in space and time. Note that, while scheme (7.26) reduces to the standard LDA scheme at steady-state, the same is not true for the ST-LDA scheme. Concerning their energy stability, the 1D analogy used in §5.4.1.1 no longer applies to the LDA scheme (7.26). While the spatial discretization certainly still benefits from the \mathcal{MU} in space, the effect of the mass-matrix originating from the distribution of the integral of the time derivative, is not understood. Conversely, for the ST-LDA scheme one might think of using its analogy with the first-order upwind scheme along the characteristic line crossing the prism. The details of this analysis are, however, unclear. For both schemes we limit ourselves to remark that certainly a degree of dissipation related to their upwind character is present, as confirmed by the fact that, when solving the system (7.16) with an iterative method, good convergence is observed. Lastly, the extension to the inhomogeneous case is simply obtained by replacing ϕ^h by the residual given in (7.20).

7.1.4.2 N schemes

A consistent extension of the N scheme to the space-time framework is of key importance for the construction of nonlinear limited schemes. As stated by proposition 7.1.6 a simple way to achieve this goal is to take scheme (5.43), combined with trapezium time integration for (5.5). Hence, we will refer to the N scheme, as to the one defined

by the space-time local nodal residual

$$\phi_i^N = \frac{|E|}{3}(u_i^{n+1} - u_i^n) + \frac{\Delta t}{2}k_i^+(u_i^n - u_{in}^n) + \frac{\Delta t}{2}k_i^+(u_i^{n+1} - u_{in}^{n+1}) \quad (7.29)$$

This is the positive first-order space-time N scheme as proposed in [7, 8]. From the analysis of chapter 5 we know that this scheme is positive if (see equation (5.44))

$$\Delta t = t^{n+1} - t^n \leq \min_{i \in \mathcal{T}_h} \frac{2|S_i|}{\sum_{E \in \mathcal{D}_i} k_i^+} \quad (7.30)$$

while its local positivity is characterized precisely by the past-shield condition (7.25). We recall that the energy stability of this scheme is characterized by propositions 4.2.2, 4.2.7 and 4.2.8. In particular, it is unconditionally energy stable. As the LDA scheme, the N scheme is \mathcal{MU} but not space-time- \mathcal{MU} . A scheme with this property, the ST-N scheme, is instead defined by

$$\phi_i^{\text{ST-N}} = \tilde{k}_i^+(u_i^{n+1} - \tilde{u}_{in}) \quad (7.31)$$

with \tilde{u}_{in} as in (7.21). The satisfaction of the past-shield condition guarantees that the ST-N scheme (7.31) satisfies the consistency condition (7.13). Moreover, it has, as the scheme defined by (5.43), a sub-element LED character, in space-time, which formally ensures the satisfaction of the local space-time discrete maximum principle (4.7). As the N and the LDA schemes, the ST-N and ST-LDA schemes are linked by

$$\phi_i^{\text{ST-N}} = \phi_i^{\text{ST-LDA}} + d_i^{\text{ST-N}} \quad (7.32)$$

where $d_i^{\text{ST-N}}$ is a space-time dissipation term given by

$$d_i^{\text{ST-N}} = \tilde{k}_i^+(u_i^{n+1} - \tilde{u}_{out}) = \sum_{j \in E} \tilde{k}_i^+ \tilde{N} \tilde{k}_j^+(u_i^{n+1} - u_j^{n+1}) \quad (7.33)$$

The space-time nature of this term is such that the ST-N schemes is generally extremely more dissipative than scheme (7.29). This will be confirmed by all our numerical results. Note that, while the N scheme of [7, 8, 118] reduces to the standard N scheme at steady-state, the same is not true for the ST-N scheme. Lastly, the extension to the non-homogeneous case is obtained as in §5.4.4 for the N scheme, while for the ST-N scheme we use (7.32) replacing ϕ^h by (7.20) in the definition of $\phi_i^{\text{ST-LDA}}$. The positivity of the LDA distribution coefficients guarantee that the hypotheses of propositions 4.3.1 and 4.3.2 and of theorem 4.3.3 are verified.

7.1.4.3 Limited schemes

Linearity preserving nonlinear schemes are obtained by applying mapping (5.65) either to the N scheme (7.29) or to the ST-N scheme (7.31). We refer to these schemes as to the limited N scheme (LN scheme) and to the limited ST-N scheme (LST-N scheme). While being \mathcal{LP} by construction, these schemes *inherit* positivity from the linear schemes due to the fact that

$$\phi_i^{\text{limited}} = \gamma_i \phi_i^{\text{linear}}, \quad \gamma_i \geq 0$$

Strictly speaking, however, last relation makes sense only if the linear scheme is locally positive. This means that, even though the positivity of the N scheme would allow the use of (7.30), when limiting the N scheme the time-step should satisfy (7.25). Additionally, we remark that the stability of these nonlinear schemes, in terms of energy dissipation, is even less clear than in the steady case. In the scalar case, however, when solving the algebraic system (7.16) with an iterative method, very fast convergence is experimented. This is a good hint of the presence of a dissipation mechanism.

7.1.5 Digression: two-layer schemes

In the previous sections we have constructed positive nonlinear \mathcal{LP} space-time schemes. For different reasons, these schemes, the LN and the LST-N schemes, are subject to the time-step limitation given by (7.25). This is particularly disappointing, considering that the schemes are by construction implicit in time. For completeness we report here a *two-layer* formulation which allows to overcome this limitation. The approach is based on an idea originally proposed in [47, 53] and later extended to the framework described here in [118, 8] and [51]. The key of the approach is to solve (7.1) at once in two space-time slabs $\Omega \times [t^n, t^{n+1}]$ and $\Omega \times [t^{n+1}, t^{n+2}]$ (see figure 7.6). The way in which this allows to use arbitrary time-steps depends on the origin of (7.25). One common cause is however the requirement that the whole discretization must ultimately lead to a time-marching scheme. The basic idea is then to add a second row of prismatic elements, in which one is free to break this condition (time-marching). The global time-marching character of the discretization procedure is then guaranteed by the satisfaction of (7.25) in the first layer, ultimately decoupling the nodal values of u^{n+2} from the ones of u^n . This gives freedom in the choice of the magnitude of $t^{n+2} - t^{n+1}$, so that the global magnitude of the time-step used in one single *time iteration*, given by $t^{n+2} - t^n$, can be arbitrarily large.

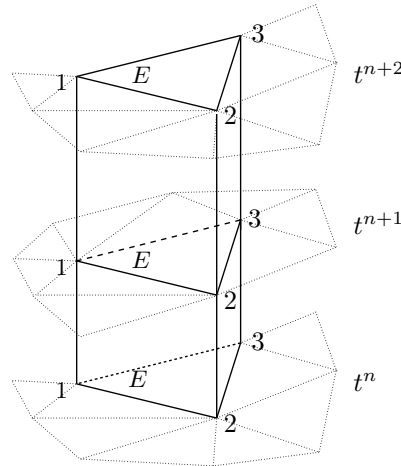


Figure 7.6: Double layer of space-time prisms

In particular, with the notation of §7.1.4, this two-layer formulation reads

$$\begin{aligned} \sum_{E \in \mathcal{D}_i} \phi_i^{n+1,n+1} + \sum_{E \in \mathcal{D}_i} \phi_i^{n+1,n+2} &= 0 & \forall i \in \mathcal{T}_h \quad \text{and} \quad t = t^{n+1} \\ \sum_{E \in \mathcal{D}_i} \phi_i^{n+2,n+2} &= 0 & \forall i \in \mathcal{T}_h \quad \text{and} \quad t = t^{n+2} \end{aligned}$$

with the \mathcal{RD} consistency requirement that $\forall E \in \mathcal{T}_h$

$$\sum_{j \in E} \phi_j^{n+1,n+1} = \phi^{n+1}, \quad \sum_{j \in E} (\phi_j^{n+1,n+2} + \phi_j^{n+2,n+2}) = \phi^{n+2}$$

To show how this works in practice, we report the two-layer N scheme of [118, 8]:

$$\begin{aligned} \phi_i^{n+1,n+1} &= \frac{|E|}{3} (u_i^{n+1} - u_i^n) + \frac{\Delta t_1}{2} k_i^+ (u_i^n - u_{in}^n) + \frac{\Delta t_1}{2} k_i^+ (u_i^{n+1} - u_{in}^{n+1}) \\ \phi_i^{n,n+1} &= 0 \\ \phi_i^{n+2,n+2} &= \frac{|E|}{3} (u_i^{n+2} - u_i^{n+1}) + \frac{\Delta t_2}{2} k_i^+ (u_i^{n+1} - u_{in}^{n+1}) \\ \phi_i^{n+1,n+2} &= \frac{\Delta t_2}{2} k_i^+ (u_i^{n+1} - u_{in}^{n+1}) \end{aligned}$$

with $\Delta t_1 = t^{n+1} - t^n$ and $\Delta t_2 = t^{n+2} - t^{n+1}$. One easily checks that the distribution adopted in the second layer, given by $\phi_i^{n+2,n+2}$ and $\phi_i^{n+1,n+2}$, defines an unconditionally positive scheme [118, 8]. Thus, Δt_2 can be arbitrarily chosen, so that $\Delta t = \Delta t_1 + \Delta t_2$ can be arbitrarily large. Application of the limiting procedure leads to a positive nonlinear \mathcal{LP} scheme which allows to perform simulations with arbitrarily large time-steps. In the case of the ST-N scheme one has, instead

$$\begin{aligned} \phi_i^{n+1,n+1} &= \tilde{k}_i^+(\Delta t_1)(u_i^{n+1} - \tilde{u}_{in}(\Delta t_1)) \\ \phi_i^{n,n+1} &= 0 \\ \phi_i^{n+2,n+2} &= \tilde{k}_i^+(\Delta t_2)(u_i^{n+2} - \tilde{u}_{in}(\Delta t_2)) \\ \phi_i^{n+1,n+2} &= \hat{k}_i^+(\Delta t_2)(u_i^{n+1} - \tilde{u}_{in}(\Delta t_2)) \end{aligned} \tag{7.34}$$

where the dependence of the space-time upwind parameters on the time-step has been added. Similarly for the inflow states, that now have to be computed using (7.21)–(7.23). In both layers the scheme is locally positive and, since Δt_1 satisfies (7.25), it is also consistent. Hence, (7.34) represents an unconditionally locally positive version of the ST-N scheme. The application of the limiting procedure leads to an unconditionally positive \mathcal{LP} nonlinear scheme. As we will remark several times in the following, the interest of this thesis is primarily the development of non-oscillatory \mathcal{LP} schemes for general \mathcal{CL} s. As a consequence, efficiency issues are often kept in the *future developments drawer*. For this reason, most of the results we will present are obtained with the single-layer formulation of the schemes. Some results computed with a conservative variant of the scheme obtained by limiting (7.34) will however be shown in the chapter devoted to the solution of the Euler equations for a perfect gas.

7.2 Nonlinear conservation laws

We consider now the solution of

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathcal{F}(u) = \mathcal{S}(x, y) \quad \text{on } \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (7.35)$$

where $\mathcal{F}(u)$ is a given nonlinear flux function. We are interested in the approximation of time-dependent solutions of (7.35) on discretizations of Ω_T which, as described in §3.1, can be decomposed into space-time slabs $\Omega \times [t^n, t^{n+1}]$ in which Ω is discretized by means of an unstructured grid composed of triangular elements, \mathcal{T}_h . The extension of space-time \mathcal{RD} schemes to the solution of (7.35) can be achieved with the same criteria discussed in chapter 6. In particular, the schemes proposed in this thesis are based on a space-time \mathcal{CRD} formulation which allows to solve general \mathcal{CL} s without the need of computing expensive mean-value linearizations of the flux Jacobian. This chapter follows the initial developments reported in [141, 142].

7.2.1 Conservative space-time schemes

As seen in chapter 2, due to its nonlinear character, equation (7.35) can develop discontinuous solutions even if the initial data are smooth. These discontinuities are characterized by the satisfaction of the Rankine-Hugoniot relations (2.18). Hence, when approximating numerically time-dependent solutions of nonlinear \mathcal{CL} s, one must ensure that across discontinuities a discrete analog of the time-dependent Rankine-Hugoniot jump conditions is verified. In the general non-homogeneous case $\mathcal{S} = 0$, this is achieved by considering the space-time schemes characterized by the following definition.

Definition 7.2.1 (Conservative space-time \mathcal{RD} scheme). *A space-time \mathcal{RD} scheme is conservative if there exist a continuous space-time approximation of the unknown u^h , of the flux \mathcal{F}^h and of the source term \mathcal{S}_h , such that*

$$\phi^h = \int_E (u^h(t^{n+1}) - u^h(t^n)) dx dy + \int_{t^n}^{t^{n+1}} \oint_{\partial E} \mathcal{F}^h \cdot \hat{n} dl dt - \int_{t^n}^{t^{n+1}} \int_E \mathcal{S}_h dx dy dt \quad (7.36)$$

Note that, in the homogeneous case $\mathcal{S} = 0$, by construction, the schemes defined in this way give a consistent approximation of the local space-time conservation law form of the problem. Thus, they ensure that a discrete analog of (2.18) is respected. As observed in chapter 6, the application of \mathcal{RD} schemes to solve (7.35) could be achieved by locally linearizing its quasi-linear analog

$$\frac{\partial u}{\partial t} + \vec{a}(u) \cdot \nabla u = 0, \quad \vec{a}(u) = \frac{\partial \mathcal{F}(u)}{\partial u}$$

In the case of the space-time schemes, using the linear approximation of u^h given by

(3.9), a conservative formulation would require the residual to be computed as

$$\begin{aligned} \phi^h = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u^h}{\partial t} + \bar{a}(u^h) \cdot \nabla u^h \right) dx dy dt = \\ \sum_{j \in E} \frac{|E|}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{j \in E} \bar{k}_j^{n+1} u_j^{n+1} + \frac{\Delta t}{2} \sum_{j \in E} \bar{k}_j^n u_j^n \end{aligned}$$

where, the \bar{k}_j^{n+1} and \bar{k}_j^n parameters are still defined as in (3.27), except that they are computed using the mean-value Jacobians

$$\bar{a}^{n+1} = \frac{2}{|E|\Delta t} \int_{t^n}^{t^{n+1}} \int_E \frac{t - t^n}{\Delta t} \bar{a}(u^h) dx dy dt$$

and

$$\bar{a}^n = \frac{2}{|E|\Delta t} \int_{t^n}^{t^{n+1}} \int_E \frac{t^{n+1} - t}{\Delta t} \bar{a}(u^h) dx dy dt$$

The schemes obtained in this way verify definition 7.2.1 with the choice $\mathcal{F}^h = \mathcal{F}(u^h)$. This approach would lead to a straightforward extension of the schemes to the nonlinear case. Moreover, the mean-value Jacobians could be replaced by approximate mean-values computed using the \mathcal{QRD} formulation of §6.2.1, and the formulas could be slightly simplified by introducing the linear in time approximation of the flux \mathcal{F}^h given by (3.14). Even so, the need of computing conservative mean-value Jacobians with sufficient accuracy leads to a considerable computational cost that, when going to systems, becomes a *bottleneck* for the efficiency of the method. Clearly, things would be much easier if the residual could be computed by approximating directly (7.36), for example with linear approximations u^h and \mathcal{F}^h . As in the steady case, we have arrived to an incompatibility between the use of the conservation law form of the problem, needed to guarantee (7.36), and the need of using the flux Jacobians in the distribution of the residual. This thesis proposes a solution to this problem based on the extension to the space-time framework of the \mathcal{CRD} technique described in §6.2.2.

7.2.2 \mathcal{CRD} for time-dependent \mathcal{CLs}

As in §6.2.2, the first element we introduce is the definition of the space-time residual. Given the continuous approximations u^h and \mathcal{F}^h of equations (3.9) and (3.14), on $E \in \mathcal{T}_h$ we compute ϕ^h by approximating (7.36) as (see equation (6.6))

$$\phi^h = \sum_{j \in E} \frac{|E|}{3} (u_j^{n+1} - u_j^n) + \Delta t \sum_{l_j=1}^3 \mathcal{F}^{l_j} \cdot \bar{n}_{l_j} - \Delta t \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (7.37)$$

where l_1 , l_2 and l_3 are the edges of E , \vec{n}_{l_j} is the exterior normal to l_j , scaled by the length of the edge and now

$$\mathcal{F}^{l_j} = \frac{1}{2} \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(u^n(x_p, y_p)) + \frac{1}{2} \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(u^{n+1}(x_p, y_p)), \quad (x_p, y_p) \in l_j \quad (7.38)$$

with $u^n(x, y)$ and $u^{n+1}(x, y)$ as in (3.9)-(3.7). For any given quadrature formula which is at least exact for a linear variation in space of \mathcal{F} , the residual given by (7.37)-(7.38) yields a consistent approximation of the integral conservation law form of (7.35) over the space-time region $E \times [t^n, t^{n+1}]$. Its accuracy is determined by the choice of NC, however the accuracy condition $\phi^h = \mathcal{O}(h^3, \Delta t^2)$ is already fulfilled by an exact integration assuming a linear variation of the flux in space. This guarantees that conservative \mathcal{LP} schemes are still second-order accurate in space and time. Clearly (7.37)-(7.38) alone does not give an approximation of (7.35) in the nodes of the grid. A distribution strategy has to be formulated.

7.2.2.1 \mathcal{LP} discretizations: \mathcal{CRD} LDA and ST-LDA schemes

The extension of linearity preserving space-time schemes to this conservative formulation is as simple as in the steady case. In fact, given \vec{a}_E , *any* linearization of the flux Jacobian over $E \times [t^n, t^{n+1}]$, and the corresponding upwind and space-time upwind parameters (see equations (3.27), (3.28) and (3.29)), the distribution coefficients of a \mathcal{LP} schemes will satisfy by construction

$$\sum_{j \in E} \beta_j = 1$$

ultimately defining a consistent splitting of (7.37)-(7.38). In particular, the schemes obtained by using the distribution coefficients given by (7.26) and (7.27), will be referred to as the \mathcal{CRD} LDA scheme and the \mathcal{CRD} ST-LDA scheme respectively.

7.2.2.2 “Monotone” discretizations: \mathcal{CRD} N and ST-N schemes

The elements introduced in §6.2.2.2 and §7.1.4.2 allow to easily describe the \mathcal{CRD} formulation of the N scheme and of the ST-N scheme. In particular, we know from §7.1.4.2 that the N scheme (7.29) corresponds to the application of the scheme (5.43) combined with trapezium time integration. The analysis of §6.2.2.2 has instead shown that the spatial discretization of the *steady-state* \mathcal{CRD} N scheme corresponds to a distribution of the balance over the element of the flux with the LDA distribution coefficient plus the addition of the cross-wind dissipation (6.11). The two information can be combined to extend the scheme of [7, 118, 8] to this conservative space-time

framework. In particular, introducing the quantities (see equation (7.38))

$$\begin{aligned}\mathcal{F}^n &= \sum_{l_j=1}^3 \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(u^n(x_p, y_p)) \cdot \vec{n}_{l_j} - \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \\ \mathcal{F}^{n+1} &= \sum_{l_j=1}^3 \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(u^{n+1}(x_p, y_p)) \cdot \vec{n}_{l_j} - \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j\end{aligned}\quad (7.39)$$

and

$$d_i^n = \sum_{j \in E} k_i^+ N k_j^+ (u_i^n - u_j^n), \quad d_i^{n+1} = \sum_{j \in E} k_i^+ N k_j^+ (u_i^{n+1} - u_j^{n+1}) \quad (7.40)$$

in unsteady computations, we will refer to the \mathcal{CRD} N scheme as to the one defined by

$$\phi_i^{\text{N-}\mathcal{CRD}} = \frac{|E|}{3} (u_i^{n+1} - u_i^n) + \frac{\Delta t}{2} \beta_i^{\text{LDA}} \mathcal{F}^n + \frac{\Delta t}{2} \beta_i^{\text{LDA}} \mathcal{F}^{n+1} + \frac{\Delta t}{2} (d_i^n + d_i^{n+1}) \quad (7.41)$$

where β_i^{LDA} is given by (7.26). Note that in general it is possible to use different averages of the flux Jacobian to compute the β_i^{LDA} used to distribute \mathcal{F}^n and \mathcal{F}^{n+1} . The same goes for d_i^n and d_i^{n+1} . In any case, the consistency of the distribution defined by (7.39), (7.40) and (7.41) with the residual given by (7.37)-(7.38) can be easily checked. The \mathcal{CRD} formulation of the ST-N scheme is obtained by using (7.32). In particular, we refer to the \mathcal{CRD} ST-N scheme, as to the one defined by

$$\phi_i^{\text{ST-N-}\mathcal{CRD}} = \beta_i^{\text{ST-LDA}} \phi^h + d_i^{\text{ST-N}} \quad (7.42)$$

where $\beta_i^{\text{ST-LDA}}$ is given by (7.27), ϕ^h is the residual computed as in (7.37)-(7.38), and $d_i^{\text{ST-N}}$ is the space-time dissipation (7.33). As observed in §7.1.4.2, the satisfaction of the past-shield condition (7.25) guarantees that the distribution defined in this way is consistent. As the title of this section says, the conservative space-time variants of the N scheme introduced lead to “monotone” approximations of weak solutions of \mathcal{CL} s. As the analysis of §6.2.2.2 shows, we cannot give a formal characterization of what we call monotonicity. We will abuse of this terminology to indicate that *in practice the schemes yield non-oscillatory approximations of weak solutions of nonlinear \mathcal{CL} s*. This will be proved on a wide variety of computational experiments.

7.2.2.3 Nonlinear schemes

If proving the robustness of the space-time \mathcal{CRD} N schemes is one of our objectives, an even more important task is to show that they can be used to produce nonlinear \mathcal{LP} schemes which are also monotone (in the sense described above). As in §6.4 and §7.1.4.3, the nonlinear schemes we will consider are obtained by applying to the space-time N schemes the mapping defined by (5.65). Due to the conservative character of the linear schemes, the sufficient condition for the well-posedness of the mapping (equation (5.68)) is always verified. So we can always define schemes obtained by limiting the \mathcal{CRD} N scheme and the \mathcal{CRD} ST-N scheme, which we refer to as the limited \mathcal{CRD} N scheme (\mathcal{CRD} LN scheme) and the limited \mathcal{CRD} ST-N scheme (\mathcal{CRD} LST-N scheme).

The lack of a formal proof of the positivity of the linear schemes, as well as the negative result of proposition 6.2.3, do not give any element to show that the nonlinear schemes we propose have any positivity property. However, the large number of numerical examples contained in the thesis will give enough evidence of their extreme robustness.

7.3 Computational examples

As in the previous chapters, we use some computational examples to confirm the theoretical constructions performed. In this chapter we present a grid convergence study for unsteady linear advection, and the approximation of a time-dependent discontinuous solution of a nonlinear \mathcal{CL} with an exponential flux. In both cases, the algebraic system arising from the space-time discretization (equation (7.16)) is solved with the explicit iterative procedure:

$$u_i^{n+1,k+1} = u_i^{n+1,k} - \frac{\delta\tau_i}{|S_i|} \sum_{E \in \mathcal{D}_i} \phi_i(u_i^{n+1,k})$$

where the iteration parameter $\delta\tau_i$ is computed as

$$\delta\tau_i = \sum_{E \in \mathcal{D}_i} \left(\frac{\Delta t k_i^+}{2} + \frac{|E|}{3} \right)$$

The time-step has been chosen such that condition (7.25) is verified:

$$\Delta t = t^{n+1} - t^n = 0.9 \min_{E \in \mathcal{T}_h} \min_{j \in E} \frac{2|E|}{3k_j^+}$$

7.3.1 Accuracy study for linear advection

We consider the solution of the linear advection equation

$$\frac{\partial u}{\partial t} + (1, 0) \cdot \nabla u = 0 \quad \text{on} \quad \Omega_T = \Omega \times [0, t_f] = ([0, 2] \times [0, 1]) \times [0, 1]$$

with the initial solution given by

$$u_0(x, y) = \begin{cases} \cos^2(2\pi r) & \text{if } r \leq 0.25 \\ 0 & \text{otherwise} \end{cases}, \quad r = \sqrt{(x - 0.5)^2 + (y - 0.5)^2}$$

From §2.1 we know that the solution is constant along characteristic lines. The initial profile is then advected in the domain, and at time $t_f = 1$ the exact solution is given by $u = u_0(x - 1, y)$. Our primary concern is to verify that the \mathcal{LP} space-time schemes indeed show second-order of accuracy. We perform a grid convergence study on a sequence of four irregular grids with mesh sizes $1/20$, $1/40$, $1/60$ and $1/80$. The problem

has been solved with the LDA, ST-LDA, LN and LST-N schemes. The accuracy is measured using the L^2 norm of the error in space

$$\|\epsilon\|_{L^2} = \sqrt{\frac{1}{n_{\text{tot}}} \sum_{i \in T_h} (u_i(t=1) - u_0(x_i - 1, y_i))^2}$$

with n_{tot} the total number of nodes. On the left on figure 7.7, we plot the \log_{10} of $\|\epsilon\|_{L^2}$ versus the logarithm of h . On the right, we report the measured order of accuracy for each successive refinement.

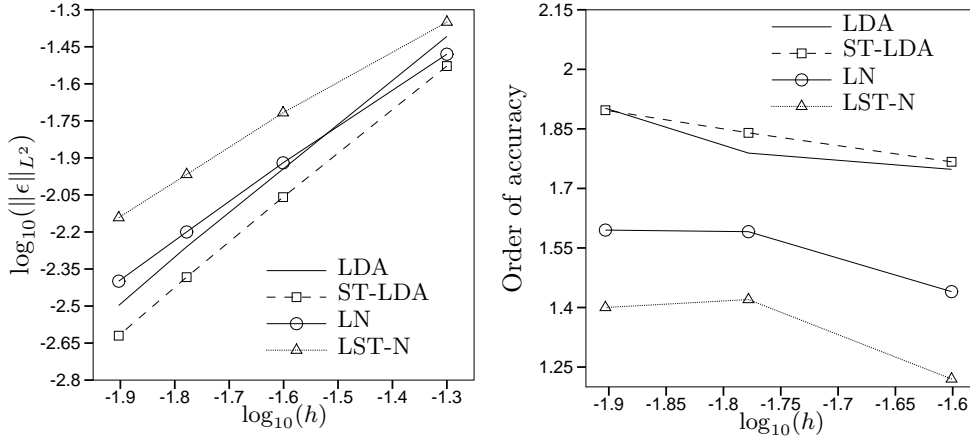


Figure 7.7: Unsteady linear advection: grid convergence. L^2 norm of the error (left) and orders of accuracy (right)

The plot on the left picture allows to make the following observations. The curves corresponding to the LDA scheme (solid line) and to the ST-LDA scheme (dashed line with squares) have roughly the same slope. The ST-LDA scheme yields, on all meshes, the smallest error. The curve corresponding to the LN scheme (solid line with circles) has a smaller slope, however the magnitude of the errors is roughly comparable with the one given by the LDA scheme. The curve corresponding to the LST-N scheme is the one with the smallest slope and largest errors. The measured orders of accuracy reported on the right picture are somehow within the *standard* and expected values for the LDA, ST-LDA and LN scheme, while the accuracy of the LST-N scheme is slightly below the expected values. In fact, both the LDA schemes give orders of accuracy between 1.75 and 1.95, which is what is normally measured on an irregular grid for a linear \mathcal{LP} scheme [12, 129, 126, 118]. Similarly, the measured accuracy for the LN scheme is between 1.45 and 1.625, which, on irregular grids, is roughly what one would expect from nonlinear \mathcal{LP} schemes [12, 129, 126, 118]. The LST-N scheme, instead, shows an accuracy which is between 1.2 and 1.4 which indeed is less than what limited \mathcal{RD} schemes normally give, even on irregular grids. Moreover, the magnitude of the errors for the LST-N scheme are markedly larger than the ones of the LN scheme. The origin of this behavior is not fully understood. One element which, we fear, determines this lack of accuracy could be a weak instability related

to energy destabilizing mechanism discussed in §5.5.2.3. We have however no formal evidence of this. In this case, we would expect a convergence rate closer to one or less. Additionally, the iterative convergence of the scheme is as good as the one of the other schemes: machine accuracy is always reached without any problem. This shows that, as observed more than once, the stability of the limited schemes is not fully understood and needs to be the subject of future study. Roughly similar conclusions have been obtained by looking at the L^1 and L^∞ norm of the error.

A *visual* characterization of the accuracy of the schemes is obtained by plotting the solutions at the final time. This is done in figures 7.8 to 7.13. For all the schemes described in this chapter we show the results obtained on an irregular grid with $h = 1/60$. In the figures, on the left we report a contour plot of the solution, and on the right, we compare the data extracted on the line $y = 0.5$ with the exact solution. To allow a better comparison, in all the left pictures the same contour levels are plotted (20 levels, from 0.01 to 1).

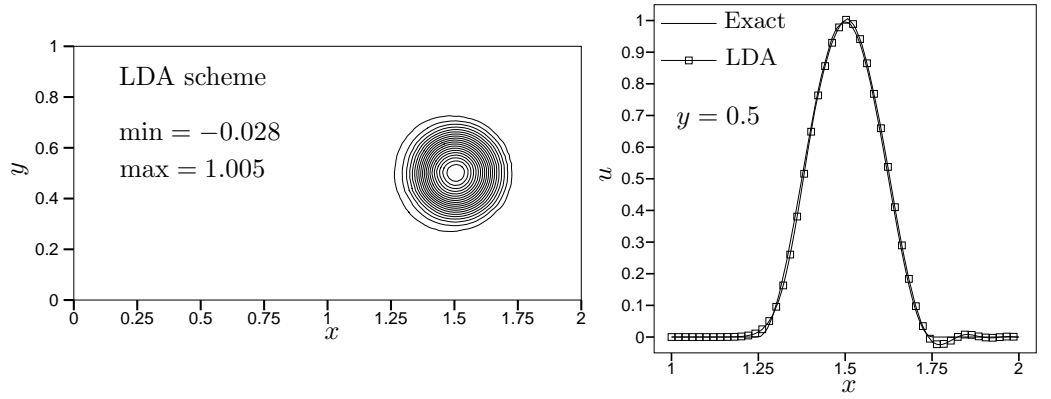


Figure 7.8: Unsteady linear advection: solution of the LDA scheme. Contour plot (left) and cut at $y = 0.5$ (right)

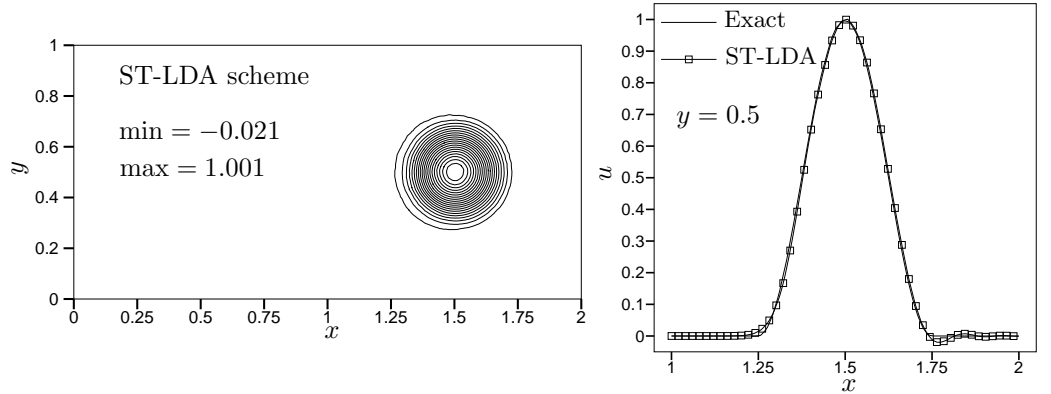


Figure 7.9: Unsteady linear advection: solution of the ST-LDA scheme. Contour plot (left) and cut at $y = 0.5$ (right)

The results of the LDA and ST-LDA schemes are shown in figures 7.8 and 7.9. Both numerical solutions are very close to the exact one. The initial peak $u_0^{\max} = 1$ is preserved and no smearing of the initial profile can be seen. This is confirmed by the plots on the right, where we also see that small oscillations appear, and that negative values are reached. This is expected, since the schemes are non-positive.

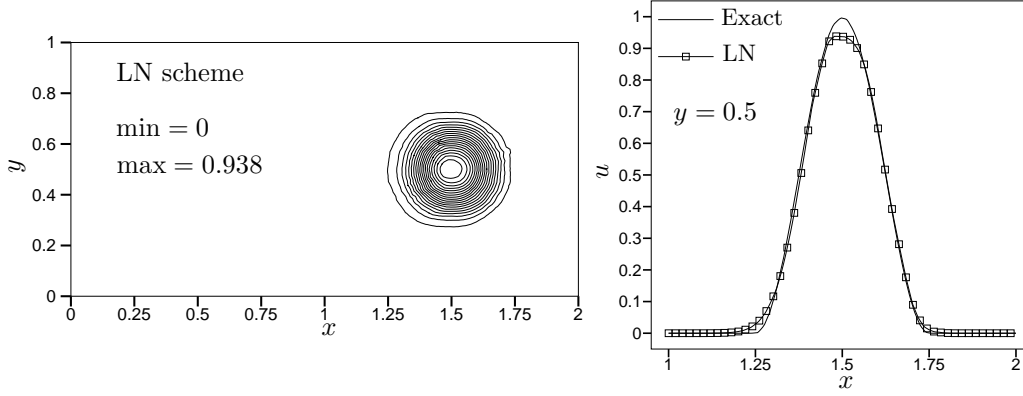


Figure 7.10: Unsteady linear advection: solution of the LN scheme. Contour plot (left) and cut at $y = 0.5$ (right)

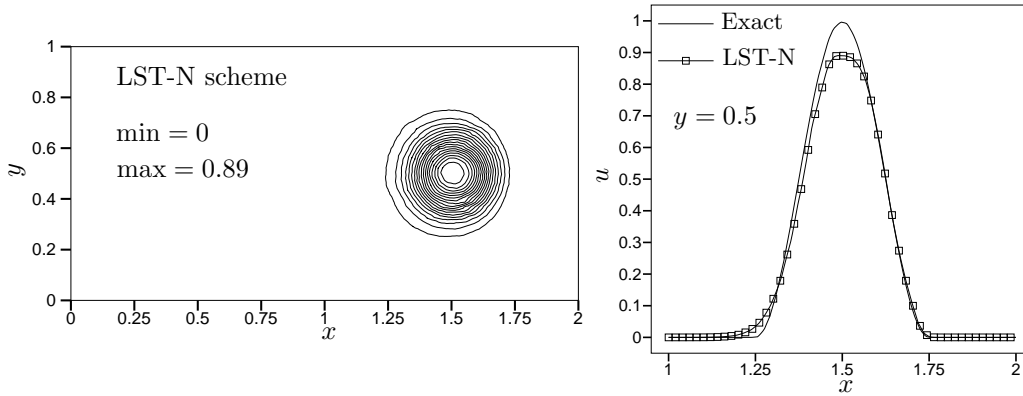


Figure 7.11: Unsteady linear advection: solution of the LST-N scheme. Contour plot (left) and cut at $y = 0.5$ (right)

The results of the LN and LST-N schemes are reported in figures 7.10 and 7.11. Both schemes keep the solution above zero, thanks to their positive character. Moreover, in both cases the exact profile is well preserved with little smearing. However, the schemes are unable to preserve the initial peak. It is worth noting that in the result obtained with the LN scheme the maximum value of the solution is 0.938, which is considerably closer to the exact value than the 0.89, given by the LST-N scheme. The line plots, on the right in the figures, confirm that the LN scheme gives a better approximation of the solution. These observations confirm the results of the grid convergence study.

Lastly, we consider the results obtained with the linear N and ST-N schemes, reported on figure 7.12 and 7.13, respectively. In both cases, the solution is kept positive but at the price of a considerable smearing of the initial profile. The maximum value of the solution is well below 1. This is not surprising, since the schemes are not \mathcal{LP} , hence at best first-order accurate. What is interesting to note is that the ST-N scheme shows a much more dissipative character than the N scheme. This is clearly seen by the contour plots of the solution. The ST-N scheme almost completely dissipates the initial profile, as shown from the fact that very few contour lines are present on the left picture on figure 7.13. Recall that in all the plots, these lines indicate the position of 20 fixed and equally spaced values of the solution, between 0.01 and 1. Only a few levels close to 0.01 can be seen in the contour plot relative to the ST-N scheme, indicating that most of the initial information is lost. Similarly, the maximum value of the solution is 0.262, which is less than half the maximum in the solution of the N scheme.

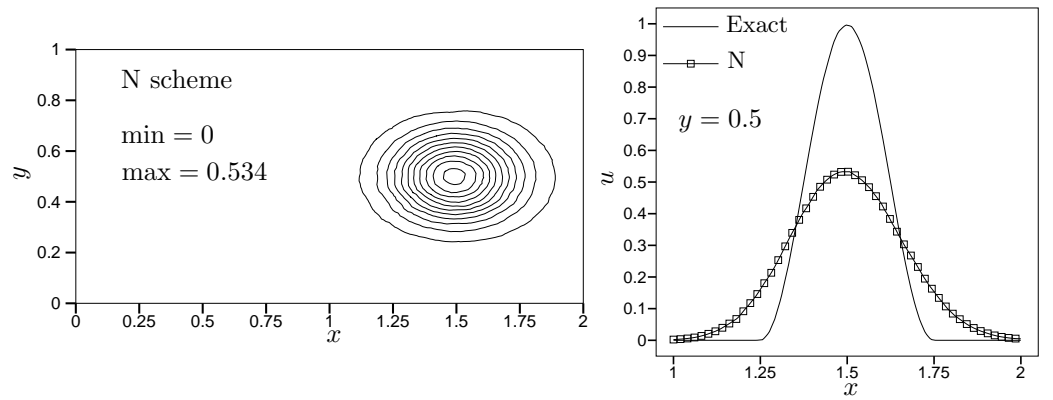


Figure 7.12: Unsteady linear advection: solution of the N scheme. Contour plot (left) and cut at $y = 0.5$ (right)

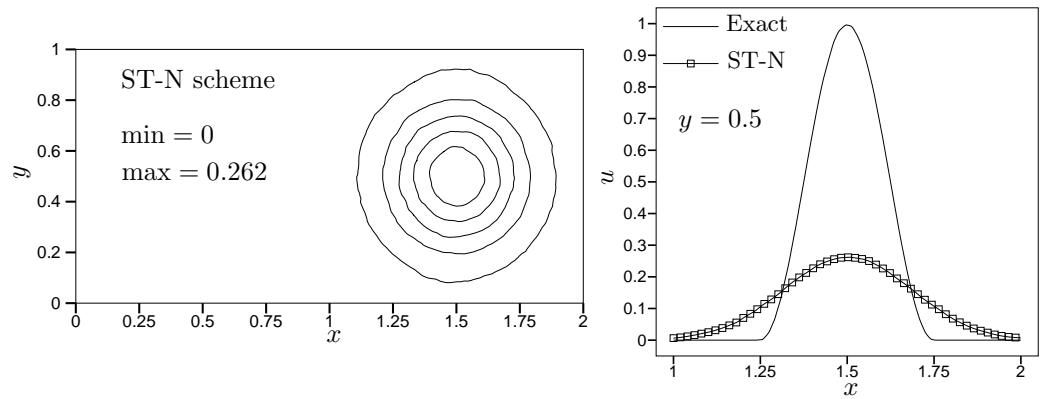


Figure 7.13: Unsteady linear advection: solution of the ST-N scheme. Contour plot (left) and cut at $y = 0.5$ (right)

7.3.2 A nonlinear problem

We consider now the solution of

$$\frac{\partial u}{\partial t} + \nabla \cdot (e^u, 0) = 0 \quad \text{on} \quad \Omega_T = \Omega \times [0, t_f] = ([-0.025, 1] \times [0, 0.08]) \times [0, 0.5]$$

with the initial solution

$$u_0(x, y) = \begin{cases} \sin(2\pi x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

In 2D space-time, this problem is identical to the 2D (space) test considered in §6.1 and §6.4. Here, the spatial domain $[-0.025, 1] \times [0, 0.08]$ is discretized with an irregular grid with $h = 1/100$. Note that the problem is basically one-dimensional, however, it can be solved on a 2D mesh. A quasi-1D solution is obtained by setting periodic boundary conditions on the upper and lower boundary. Due to the constant spacing of the nodes along the boundaries, this condition has been imposed in a strong nodal sense, as described in [99]. A reference solution at time $t_f = 0.5$ has been computed in 1D, solving the problem with the limited Lax-Wendroff scheme of [159] on 5000 1D cells. When applying the space-time \mathcal{CRD} schemes described in this chapter we have used trapezium rule in (7.38), while the flux Jacobians have been linearized as in §6.4:

$$\vec{a}_E = \frac{1}{3} \sum_{j \in E} \vec{a}(u_j^{n+1})$$

Probably due to the formation of a moving shock in the early times of the computation, we were not able to obtain a solution with the \mathcal{LP} LDA and ST-LDA schemes. We present instead the results obtained with the \mathcal{CRD} N and \mathcal{CRD} ST-N schemes, and with their limited variants, the \mathcal{CRD} LN and \mathcal{CRD} LST-N schemes. We compare with the reference solution the data extracted on the line $y = 0.04$.

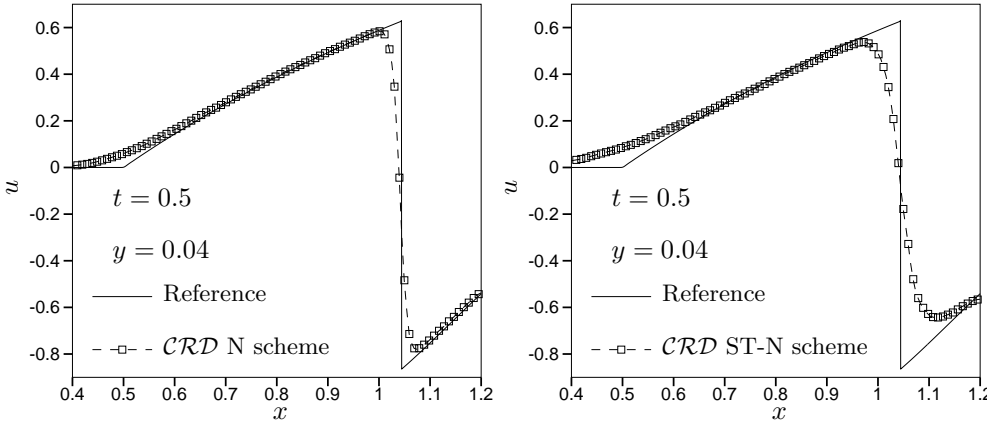


Figure 7.14: Nonlinear \mathcal{CL} with *exponential flux*. Solution at time $t = 0.5$. Data at $y = 0.04$. Left: \mathcal{CRD} N scheme. Right: \mathcal{CRD} ST-N scheme

Figure 7.14 shows the results obtained with the \mathcal{CRD} N (on the left) and ST-N (on the right) schemes. In both solutions the position of the shock is correct, which shows the conservative character of the discretization. Moreover, both schemes yield a monotone approximation of the discontinuity, which, however, is smeared over several cells. In particular, the ST-N scheme gives a much wider shock layer, showing, as in the previous problem, a larger numerical dissipation. This is attributed to the space-time character of the ST-N scheme. Indeed, for the N scheme only the spatial discretization introduces dissipation in the discretization, while the trapezium time integration does not have a dissipative character, or, at least, not a strongly dissipative one (see §6.3.3.3). On the other hand, for the ST-N scheme, temporal and spatial discretizations are coupled, leading to the simultaneous introduction of entropy dissipation in space and time (see equations (7.42) and (7.32)).

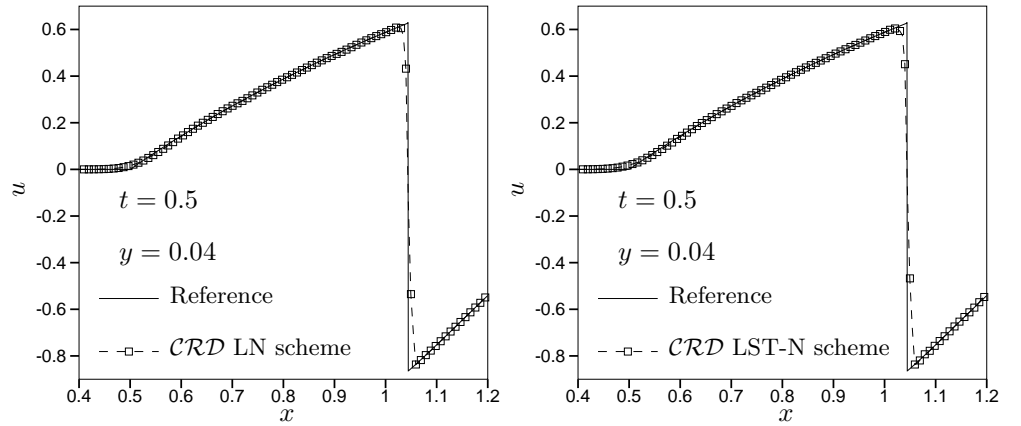


Figure 7.15: Nonlinear \mathcal{CL} with *exponential flux*. Solution at time $t = 0.5$. Data at $y = 0.04$. Left: \mathcal{CRD} LN scheme. Right: \mathcal{CRD} LST-N scheme

The results obtained with the space-time nonlinear \mathcal{CRD} LN and LST-N schemes are instead reported in figure 7.15. The difference in the two solutions is absolutely negligible, both in the smooth part and across the discontinuity. The latter is monotonically approximated and it is extremely sharp, especially considering that the computation has been performed on a 2D mesh. Also, very good iterative convergence has been observed for both schemes.

7.4 Summary

This chapter has extended the conservative schemes constructed in chapter 6 to a space-time framework. This allows to obtain second-order accurate approximation of time-dependent weak solutions of \mathcal{CL} s. The most important elements introduced are summarized hereafter.

- An improved compact cell-vertex prototype for unsteady scalar advection has been introduced and shown to overcome the accuracy limitations of the discretizations considered in the previous chapters. This is achieved by introducing a coupling of the temporal evolution of the solution in the nodes, through a mass-matrix;
- A fully discrete formulation for unsteady advection has been achieved by introducing a space-time \mathcal{RD} framework. This framework encompasses all the schemes represented by the new prototype as well as the schemes considered in the previous chapters, if the time derivative is discretized with the trapezium scheme;
- The accuracy of space-time \mathcal{RD} schemes has been studied, giving a formal condition to achieve second-order of accuracy in space and time. Linearity preserving space-time \mathcal{RD} schemes have been shown to respect this condition;
- A geometrical analysis of the space-time schemes has allowed to defined a new class of space-time- \mathcal{MU} discretizations. These schemes incorporate a complete coupling of spatial and temporal discretization. A condition ensuring the time-marching character of the whole discretization has been derived. This past-shield condition involves a time-step restriction equivalent to one ensuring the local positivity of the N scheme with \mathcal{CN} time-integration;
- Space-time variants of the LDA and of the N schemes have been presented. Both schemes admit (at least) two different formulations, one which is \mathcal{MU} only in space (referred to as the LDA and N schemes), while the other is space-time- \mathcal{MU} (referred as to the ST-LDA and ST-N schemes). In the case of the N scheme, the first formulation is simply obtained by combining the \mathcal{MU} spatial discretization of the N scheme presented in §5.4.2 with \mathcal{CN} time integration. This scheme is conditionally positive and unconditionally energy stable. The space-time dissipation introduced by the ST-N scheme has been discussed;
- Nonlinear space-time schemes are obtained by formally applying the limiting procedure of §5.5.2 to the N and ST-N schemes. We are not able to characterize the energy stability of the limited space-time schemes;
- The double layer formulation of space-time \mathcal{RD} has been recalled. This formulation allows to construct unconditionally positive nonlinear schemes;
- The space-time schemes have been extended to the solution of time-dependent nonlinear \mathcal{CL} s using the \mathcal{CRD} technique;
- Conservative variants of the space-time LDA and N schemes have been introduced. Nonlinear schemes are obtained by limiting the space-time \mathcal{CRD} schemes. As in the steady case, we are not able to prove formally the positivity of the \mathcal{CRD} N and limited N schemes;
- A grid convergence study for unsteady linear advection confirms the expected levels of accuracy for the LDA, ST-LDA and limited N schemes. The accuracy of the limited ST-N scheme is slightly less than expected. The reasons are not understood. The results on this linear problem show that the N and limited N schemes have a markedly lower error than the ST-N and limited ST-N schemes. The ST-N scheme shows a very large numerical dissipation;

- The solution of a nonlinear \mathcal{CL} with an exponential flux has been considered. The solution involves the formation of a moving shock. The space-time \mathcal{CRD} LDA and ST-LDA schemes did not give any solution, due to the presence of the shock. The space-time \mathcal{CRD} N and ST-N schemes have confirmed their conservative character. Moreover, they produced a monotonic approximation of the shock. The results obtained with the space-time \mathcal{CRD} limited N and ST-N schemes are practically perfect: monotone and sharp capturing of the moving shock. The differences between the two nonlinear schemes on the solution of this problem are absolutely negligible.

Chapter 8

Extension to systems

In Chapters 4 to 7 we have built and analyzed compact schemes for the solution of nonlinear conservation laws on unstructured meshes. Their conservative character is guaranteed by the use of the integral formulation of the problem for the definition of the local element residual. This gives a flexibility and generality which allows to use them for the solution of very nonlinear \mathcal{CL} s without having to seek for costly conservative mean-value linearizations of the flux Jacobian. While this advantage might be of little importance for scalar equations, the case of a system of \mathcal{CL} s is different.

We recall that a large number of fluid-mechanics phenomena can be modeled by the nonlinear set of PDEs

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{u}) = \mathcal{S}(x, y), \quad \text{on } \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (8.1)$$

where now the unknown \mathbf{u} and the source term \mathcal{S} are m -vectors, while in two dimensions \mathcal{F} is a $m \times 2$ -tensor, m being the number of *conserved quantities* (e.g. mass, momentum and energy). In many cases, the dependence of \mathcal{F} on \mathbf{u} is highly nonlinear. Thus, the numerical approximation of (8.1) requires the use of techniques as flexible as possible, so that a correct approximation of its weak solutions is guaranteed. The construction performed in the previous chapters has put us half-way from this objective. We have schemes able to approximate the scalar counterpart of (8.1) on unstructured meshes. They are as general and flexible as allowed by the problem, in the sense that they are only constrained by the requirement of respecting the conservative character of the model and of encompassing some kind of dissipative mechanism. This is guaranteed by some form of upwinding of the information. Moreover, some of these schemes guarantee a stable and non-oscillatory resolution of discontinuities while still being second-order accurate. These features have been proved formally and with the illustrative computational examples given in the previous chapters. This chapter shows one possible direction that can be undertaken to fill the remaining gap: the extension of residual distribution schemes to systems.

The approach used in this thesis is probably the simplest possible, since it is based on a formal extension of the schemes, making use of a *matrix formulation* [177, 178, 179]. As we will see, while being straightforward, this approach leads to a loss of understanding of the properties of the discretization, from a physical and from a formal point of view. Most of the analogies used in the scalar case to analyze the schemes still hold formally but lose their physical or geometrical meaning, while other properties are somehow lost or need to be replaced by less rigorous arguments. On the other hand, the tools introduced for the energy/entropy stability and accuracy analyses extend formally to systems, allowing to shorten the presentation. We remark that other techniques can be used to extend to systems the conservative schemes developed in this thesis. Part of the extensive literature on \mathcal{RD} schemes given in chapter 1 describes some of these different approaches, and we refer the interested reader to these references for an overview.

The chapter is divided into three main blocks plus a summary. The first part describes the extension of the schemes to the approximation of steady-state solutions of linear symmetric systems. The basics of the matrix formulation of the method are introduced here. Little additional theoretical analysis is performed. The matrix variants of the LDA, N and limited N schemes are introduced. Then, in the second part of the chapter, we explain how time-dependent solutions of linear symmetric hyperbolic systems are approximated, using a space-time matrix residual distribution technique. In this case, we limit ourselves to present the main concepts and the space-time matrix formulation of the LDA and N schemes, the limited schemes being constructed exactly as in the steady case. The third part of this chapter highlights the effectiveness of the conservative approach proposed in the thesis, showing how the schemes are applied to the solution of (8.1). In this case, very little is possible to prove formally, and one relies on numerical experiments to verify the properties of the method. These experiments will be the subject of the rest of the thesis.

8.1 Matrix \mathcal{RD} for linear symmetric systems

Consider the approximation of the steady limit of the hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + A_1 \frac{\partial \mathbf{u}}{\partial x} + A_2 \frac{\partial \mathbf{u}}{\partial y} = \mathcal{S}(x, y), \quad \text{on } \Omega \subset \mathbb{R}^2 \quad (8.2)$$

where the constant matrices A_1 and A_2 are symmetric. As observed in the beginning of chapter 2, the system being hyperbolic, $\forall \xi \in \mathbb{R}^2$ the matrix

$$K(\vec{\xi}) = A_1 \xi_1 + A_2 \xi_2 \quad (8.3)$$

admits a complete set of real eigenvalues and linearly independent eigenvectors, and can be decomposed as in equations (2.5) and (2.6). Following the definition 5.1.1 of a \mathcal{RD} scheme, given \mathcal{T}_h , the unstructured discretization of the spatial domain, solutions of (8.2) are approximated as follows [179, 178, 177]

1. $\forall E \in \mathcal{T}_h$ we compute the element residual

$$\phi^h = \int_E \left(A_1 \frac{\partial \mathbf{u}_h}{\partial x} + A_2 \frac{\partial \mathbf{u}_h}{\partial y} - \mathbf{S}_h \right) dx dy \quad (8.4)$$

where \mathbf{u}_h and \mathbf{S}_h are given by (3.7) and (3.15). Using the properties of the basis functions (3.6), the residual can be shown to be given by

$$\phi^h = \sum_{j \in E} K_j \mathbf{u}_j - \frac{|E|}{3} \sum_{j \in E} \mathbf{S}_j \quad (8.5)$$

where the K_j s are given by (3.16) and are a matrix generalization of the scalar k_j parameters in (5.7).

2. $\forall E \in \mathcal{T}_h$ distribute fractions of ϕ^h to the nodes of E . Denoting by ϕ_i the split residuals or local nodal residuals, *by construction* we must have

$$\sum_{j \in E} \phi_j = \phi^h \quad (8.6)$$

If there exist matrices β_i such that $\phi_i = \beta_i \phi^h$, then

$$\sum_{j \in E} \beta_j = \mathbf{I} \quad (8.7)$$

with \mathbf{I} the $m \times m$ identity matrix. The matrix β_i is called a *distribution matrix*.

3. $\forall i \in \mathcal{T}_h$ assemble the contributions from all $E \in \mathcal{D}_i$ and evolve \mathbf{u}_i in time

$$|S_i| \frac{d\mathbf{u}_i}{dt} + \sum_{E \in \mathcal{D}_i} \phi_i = 0 \quad (8.8)$$

The schemes defined by the above three steps are clearly a formal generalization of the \mathcal{RD} schemes of definition 5.1.1. The difference is that the unknown and the residual are now vectors, hence a matrix formalism is needed. The complexity introduced by this formalism, however, leaves some properties unchanged, as for example:

Proposition 8.1.1 (Matrix \mathcal{RD} , accuracy and \mathcal{LP} schemes). *The matrix \mathcal{RD} schemes defined by (8.4), (8.6) and (8.8) are only first-order accurate in space in time-dependent computations. However, they are second-order accurate at steady-state if*

$$\phi_i = \mathcal{O}(h^3)$$

In particular, since $\phi^h = \mathcal{O}(h^3)$, the linearity preserving schemes defined by $\phi_i = \beta_i \phi^h$, with β_i uniformly bounded with respect to h and to the data of the problem, are second-order accurate at steady-state.

The proof of this proposition is *identical* to the proof of the scalar condition for second-order of accuracy (see §4.1.1, §4.4.2, §4.2.3 and [3, 9]). Similarly, the definition of a linear scheme remains basically unchanged.

Definition 8.1.2 (Linear matrix \mathcal{RD}). A matrix \mathcal{RD} scheme is linear if

$$\phi_i = \sum_{\substack{j \in E \\ j \neq i}} C_{ij}^E (\mathbf{u}_i - \mathbf{u}_j) - |E| \sum_{j \in E} C_{ij}^{\mathcal{S}} \mathbf{S}_j$$

with matrices C_{ij}^E and $C_{ij}^{\mathcal{S}}$ independent of the numerical solution \mathbf{u}_h .

Another property admitting a formulation analog to the scalar case is the energy stability. In particular, for systems the equivalence lemma 4.2.3 is still valid and allows to characterize the dissipative character of the schemes. An example of its application will be given for the matrix N scheme. As for scalar problems, the dissipation characteristics of the matrix \mathcal{RD} schemes are related to a multidimensional upwinding property. To define this property, on $E \in \mathcal{T}_h$ we introduce, as in the scalar case, inflow and outflow states defined as

$$\mathbf{u}_{in} = \left(\sum_{j \in E} K_j^- \right)^{-1} \sum_{j \in E} K_j^- \mathbf{u}_j = - \sum_{j \in E} N K_j^- \mathbf{u}_j \quad (8.9)$$

and

$$\mathbf{u}_{out} = \left(\sum_{j \in E} K_j^+ \right)^{-1} \sum_{j \in E} K_j^+ \mathbf{u}_j = \sum_{j \in E} N K_j^+ \mathbf{u}_j \quad (8.10)$$

with K_j^\pm given by (3.18), and having introduced the matrix (see (3.17) and (3.19))

$$N = \left(\sum_{j \in E} K_j^+ \right)^{-1} = - \left(\sum_{j \in E} K_j^- \right)^{-1} = \frac{1}{2} \left(\sum_{j \in E} |K_j| \right)^{-1} \quad (8.11)$$

For the moment, we shall assume that the hyperbolicity of the linear system (8.2), guarantees that N exists always. Using this notation, one easily shows that the residual can be expressed as

$$\phi^h = \left(\sum_{j \in E} K_j^+ \right) (\mathbf{u}_{out} - \mathbf{u}_{in}) \quad (8.12)$$

Even though (8.12) is formally identical to (5.11), it is known that if system (8.2) is not *diagonalizable*, that is if A_1 and A_2 do not commute, then the relation between the inflow and outflow states and the speeds associated to the eigenvalues of each K_j is complex and does not really allow to see (8.12) as a 1D balance along precise directions in space. However, one can still define \mathcal{MU} schemes, where, for each node $i \in E$, the eigenvalues of K_i are used as a reference to decide whether i receives a larger or smaller amount of residual.

Definition 8.1.3 (\mathcal{MU} matrix \mathcal{RD} schemes). A matrix \mathcal{RD} scheme is \mathcal{MU} if

$$K_i^+ = 0 \implies \phi_i = 0$$

In particular, for \mathcal{MU} schemes one has $\phi_i \propto K_i^+$.

Due to the coupling of the equations, we are not able to define 1-target and 2-target triangles, all the elements being in general 3-target. Since the \mathcal{MU} property was particularly beneficial in the 1-target situations (see propositions 5.4.2 and 6.3.5), this

might seem an important loss. However, we will show that also for matrix \mathcal{RD} the multidimensional upwinding introduces stabilizing effects, at least from the point of view of energy dissipation.

The property which seems more difficult to analyze in the case of a system is the L^∞ stability of the schemes. Not surprisingly, this reflects the fact that also for exact solutions of (8.2), this property cannot be formulated in a rigorous way (see §2.3). We mention that, for a linear symmetric system, a criterion allowing to characterize this property has been recently proposed in [118, 10, 9]. Here we limit ourselves to recall that the technique proposed in the references uses the idea that basic solutions of (8.2) can be represented as compositions of simple waves. The authors then introduce wave decompositions of the discrete solution \mathbf{u}_h . These decompositions can be written as

$$\mathbf{u}_h = \sum_{\sigma=1}^m \varphi^\sigma(x, y) \mathbf{r}_\sigma$$

where, given a direction $\vec{\xi} \in \mathbb{R}^2$, \mathbf{r}_σ is the σ -th eigenvector of $K(\vec{\xi})$ (equation (8.3)). The idea is then that, when applying a linear scheme to \mathbf{u}_h , the local nodal residuals ϕ_i can also be expressed as a sum of waves:

$$\phi_i = \sum_{j \in E} \sum_{\sigma=1}^m c_{ij}^\sigma (\varphi_i^\sigma - \varphi_j^\sigma) \mathbf{r}_\sigma$$

where $\varphi_j^\sigma = \varphi^\sigma(x_j, y_j)$ and the coefficients c_{ij}^σ depend on the scheme. This decomposition, ultimately allows the authors of [118, 10, 9] to extend LED and positivity analyses to each simple wave, thus giving a tool to formally study the stability of the schemes.

In the following sections, we present the matrix variants of the \mathcal{MU} LDA and N schemes. We then discuss the construction of nonlinear matrix schemes for systems.

8.1.1 Matrix LDA scheme

The matrix variant of the LDA scheme is obtained simply by defining a distribution matrix formally identical to (5.39). In particular, we refer to the matrix LDA scheme as to the one defined by

$$\phi_i^{\text{LDA}} = \beta_i^{\text{LDA}} \phi^h, \quad \beta_i^{\text{LDA}} = K_i^+ N \quad (8.13)$$

The scheme is clearly linear, \mathcal{MU} and \mathcal{LP} . It has no L^∞ stability properties.

8.1.1.1 Energy production of the matrix LDA scheme

The discrete energy associated to the approximation \mathbf{u}_h is defined exactly as in the scalar case. With the notation of (4.30), we have

$$\mathcal{E}_h = \sum_{i \in \mathcal{T}_h} \frac{\mathbf{u}_i^T |S_i| \mathbf{u}_i}{2} = \frac{\mathbf{U}^T D_{|S_i|} \mathbf{U}}{2}$$

where \mathbf{U} is an array with a block structure, the i -th block entry being \mathbf{u}_i . As in the scalar case, one easily shows that the evolution of \mathcal{E}_h in time can be decomposed in a sum of elemental energy production/dissipation rates (see §4.2):

$$\frac{d\mathcal{E}_h}{dt} = \sum_{E \in \mathcal{T}_h} \frac{d\mathcal{E}}{dt}$$

In the case of the LDA scheme, one easily shows that the local energy evolution is governed by the following equation

$$\frac{d\mathcal{E}^{\text{LDA}}}{dt} = -\frac{1}{2}(\mathbf{u}_{out} + \mathbf{u}_{in})^T \left(\sum_{j \in E} K_j^+ \right) (\mathbf{u}_{out} - \mathbf{u}_{in}) - \epsilon^{\text{LDA}} \quad (8.14)$$

with

$$\epsilon^{\text{LDA}} = \frac{1}{2}(\mathbf{u}_{out} - \mathbf{u}_{in})^T \left(\sum_{j \in E} K_j^+ \right) (\mathbf{u}_{out} - \mathbf{u}_{in})$$

Note that each K_j^+ is symmetric positive semi-definite by definition. Hence $\epsilon^{\text{LDA}} \geq 0$ represents a rate of energy dissipation. However, as already remarked, in the case of matrix \mathcal{RD} , \mathbf{u}_{in} and \mathbf{u}_{out} cannot be associated to precise spatial directions, as was done in §5.1.1 and §5.4.1.1, hence the first term appearing on the right hand side in the energy balance (8.14) can hardly be interpreted. It is however noteworthy, that equation (8.14) is equivalent to the local balance of the 1D first-order CIR upwind scheme [110]. Even though it is not clear how much physical sense equation (8.14) has, it shows that the \mathcal{MU} does introduce energy dissipation in the discretization.

8.1.2 Matrix N scheme

As for the LDA scheme, the matrix the N scheme is a formal generalization of its scalar counterpart, equations (5.43), (5.48) and (5.53). In particular, we refer to the matrix N scheme as to the one defined by

$$\phi_i^{\text{N}} = K_i^+(\mathbf{u}_i - \mathbf{u}_{in}) - \beta_i^{\text{LDA}} \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (8.15)$$

with β_i^{LDA} as in (8.13). The matrix N scheme is \mathcal{MU} but is not \mathcal{LP} . However, in [10, 118, 9] it is proved that, in the homogeneous case, the matrix N scheme is L^∞ -stable on simple-wave solutions.

8.1.2.1 Energy stability

The analysis of the energy stability of the matrix N scheme has been initially reported in [18] and then in [4, 9]. As in the scalar case, the local evolution of the discrete energy \mathcal{E}_h can be written as (see equation (5.46))

$$\frac{d\mathcal{E}^N}{dt} = - \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}^T \overline{M}^N \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} - \frac{1}{2} \sum_{j \in E} \mathbf{u}_j^T K_j \mathbf{u}_j \quad (8.16)$$

where \overline{M}^N is the block matrix given by (see §4.2 and §5.4.2.1, equation (5.46))

$$\begin{aligned} \overline{M}^N = & \frac{1}{2} \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix}^T N \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix} + \\ & \frac{1}{2} \begin{bmatrix} K_1^+ & 0 & 0 \\ 0 & K_2^+ & 0 \\ 0 & 0 & K_3^+ \end{bmatrix} - \frac{1}{2} \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix}^T N \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix} + \\ & \frac{1}{2} \begin{bmatrix} -K_1^- & 0 & 0 \\ 0 & -K_2^- & 0 \\ 0 & 0 & -K_3^- \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -K_1^- \\ -K_2^- \\ -K_3^- \end{bmatrix}^T N \begin{bmatrix} -K_1^- \\ -K_2^- \\ -K_3^- \end{bmatrix} \quad (8.17) \end{aligned}$$

As in the scalar case, for a linear system the second term on the right hand side in (8.16) cancels identically when summing over all the elements of the mesh (see §5.2.1):

$$\sum_{E \in \mathcal{T}_h} \sum_{j \in E} \mathbf{u}_j^T K_j \mathbf{u}_j = \sum_{j \in \mathcal{T}_h} \sum_{E \in \mathcal{D}_j} \mathbf{u}_j^T K_j \mathbf{u}_j = \sum_{j \in \mathcal{T}_h} \mathbf{u}_j^T \left(\sum_{E \in \mathcal{D}_j} K_j \right) \mathbf{u}_j = 0$$

By virtue of the energy equivalence lemma 4.2.3, we can limit ourselves to study the properties of the *equivalent energy operator* \overline{M}^N . In [18, 4, 9] it is proved that this operator is positive semi-definite, hence *the matrix N scheme is energy stable*. With a proof identical to the scalar case, we can show that

Proposition 8.1.4 (Discrete energy stability - θ -scheme, system case). *The family of schemes represented by the θ -scheme (4.20) verify the following fully discrete energy balance*

$$\mathcal{E}_h^{n+1} = \mathcal{E}_h^n - \Delta t \left(\theta \mathbf{U}^{n+1} + (1 - \theta) \mathbf{U}^n \right)^T M^{\mathcal{E}_h} \left(\theta \mathbf{U}^{n+1} + (1 - \theta) \mathbf{U}^n \right) - (2\theta - 1) \epsilon_h \quad (8.18)$$

where $M^{\mathcal{E}_h}$ is the symmetric block matrix energy operator of the spatial discretization, and ϵ_h is the discrete energy production in time, given by

$$\epsilon_h = \frac{1}{2} (\mathbf{U}^{n+1} - \mathbf{U}^n)^T D_{|S_i|} (\mathbf{U}^{n+1} - \mathbf{U}^n) \geq 0.$$

The time discretization has a stabilizing effect for $\theta > 1/2$ and a destabilizing effect for $\theta < 1/2$. In particular, the explicit FE time discretization has the maximum energy destabilizing character and the implicit BE scheme is the most stable. The CN scheme is the only time discretization preserving the dissipation properties of the spatial discretization. For this reason the CN scheme is said to be energy conservative.

As a result of this proposition, *the matrix N scheme is unconditionally energy stable when (8.8) is discretized with the implicit BE scheme or with the CN scheme.* Lastly, we note that the N scheme can be written as

$$\phi_i^N = \beta_i^{\text{LDA}} \phi^h + \mathbf{d}_i^N, \quad \mathbf{d}_i^N = \sum_{j \in E} K_i^+ N K_j^+ (\mathbf{u}_i - \mathbf{u}_j) \quad (8.19)$$

The contribution of $\{\mathbf{d}_j^N\}_{j \in E}$ to the energy balance can be written as (see §5.4.3)

$$\epsilon^N = \sum_{j \in E} \mathbf{u}_j^T \mathbf{d}_j^N = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}^T D^N \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}$$

where D^N is the block symmetric matrix (see §5.4.3, equation (5.50))

$$D^N = \begin{bmatrix} K_1^+ & 0 & 0 \\ 0 & K_2^+ & 0 \\ 0 & 0 & K_3^+ \end{bmatrix} - \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix} N \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix}^T \quad (8.20)$$

The positive semi-definiteness of D^N is proved in [18, 4, 9]. Hence, $\{\mathbf{d}_j^N\}_{j \in E}$ define dissipation terms. In particular, the N scheme is more dissipative than the LDA scheme.

8.1.3 Nonlinear matrix \mathcal{RD} schemes

The last missing element to extend \mathcal{RD} schemes to the solution of (8.2) is the construction of \mathcal{LP} nonlinear schemes with L^∞ stability properties. The approach used here is the same proposed in [10, 118, 9]. In these references, the authors propose a wave decomposition technique to analyze the L^∞ stability of matrix \mathcal{RD} schemes. This technique, justifies the following construction of limited schemes for systems.

1. Given a direction $\vec{\xi} \in \mathbb{R}^2$, compute the left and right eigenvectors of $K(\vec{\xi})$ (equation (8.3)). Let $\{\mathbf{l}_\sigma\}_{\sigma=1}^m$ and $\{\mathbf{r}_\sigma\}_{\sigma=1}^m$ be these eigenvectors.
2. Let $\{\phi_j^\infty\}_{j \in E}$ be the local nodal residuals of a linear first-order scheme which is L^∞ -stable on simple waves. For $\sigma = 1, m$, compute scalar local nodal residuals and scalar element residuals by projecting ϕ_j^∞ and ϕ^h on $\{\mathbf{l}_\sigma\}_{\sigma=1}^m$:

$$\begin{aligned} \varphi_{j,\sigma}^\infty &= \mathbf{l}_\sigma^T \phi_j^\infty \\ \varphi_\sigma^h &= \mathbf{l}_\sigma^T \phi^h \end{aligned}$$

3. For $\sigma = 1, m$, limit each scalar component $\varphi_{j,\sigma}^\infty$ to obtain a set of nonlinear \mathcal{LP} scalar nodal residuals:

$$\varphi_{j,\sigma} = \beta_{j,\sigma} \varphi_\sigma^h$$

4. Project back the nodal residuals onto the space of conserved variables:

$$\boldsymbol{\phi}_i = \sum_{\sigma=1}^m \varphi_{i,\sigma} \mathbf{r}_\sigma$$

The scheme obtained in this way is \mathcal{LP} by construction. For a linear symmetric system, in [10, 118, 9] it is shown that it is also L^∞ -stable on simple waves. Unfortunately, differently from the scalar case (see §5.5.1.1, §5.5.1.2 and §5.5.2.3), we are absolutely unable to characterize the energy stability properties of the schemes obtained in this way. The only remark that can be made is that since for matrix \mathcal{RD} schemes, even for \mathcal{MU} schemes, real 1-target elements are generally absent or very few, the procedure described by steps 1.-4. is likely to produce the destabilizing effects described in §5.5.2.3. As we will see, this is confirmed from the fact that, in the system case, convergence to machine zero is never achieved for these schemes.

8.2 Space-time matrix \mathcal{RD}

Section §8.1 has introduced the matrix formulation of residual distribution for the approximation of the steady limit of symmetric hyperbolic systems of PDEs. Here we consider, instead, the solution of

$$\frac{\partial \mathbf{u}}{\partial t} + A_1 \frac{\partial \mathbf{u}}{\partial x} + A_2 \frac{\partial \mathbf{u}}{\partial y} = \mathcal{S}(x, y) \quad \text{on} \quad \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (8.21)$$

for $t_f > 0$ finite. We will present a matrix formulation of the space-time \mathcal{RD} schemes of §7.1.2-§7.1.5. As before, the extension is done in a formal way. Very few new ideas have to be introduced, while most of the geometrical analogies are lost. Consider then a space-time slab $\Omega \times [t^n, t^{n+1}]$. With the notation of (3.9), given \mathbf{u}^n , one computes the nodal values of \mathbf{u}^{n+1} in the three following steps

1. $\forall E \in \mathcal{T}_h$ compute on the space-time prism $E \times [t^n, t^{n+1}]$ the residual

$$\boldsymbol{\phi}^h = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial \mathbf{u}^h}{\partial t} + A_1 \frac{\partial \mathbf{u}^h}{\partial x} + A_2 \frac{\partial \mathbf{u}^h}{\partial y} - \mathcal{S}_h \right) dx dy \quad (8.22)$$

where \mathbf{u}^h is the space-time approximation (3.9) and \mathcal{S}_h is the spatial approximation of the source term (3.15). Straightforward calculations show that

$$\boldsymbol{\phi}^h = \sum_{j \in E} \frac{|E|}{3} (\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) + \frac{\Delta t}{2} \sum_{j \in E} (K_j \mathbf{u}_j^n + K_j \mathbf{u}_j^{n+1}) - \Delta t \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (8.23)$$

2. $\forall E \in \mathcal{T}_h$ distribute $\boldsymbol{\phi}^h$ to the nodes of E . Denoting by $\boldsymbol{\phi}_i$ the split residuals or local nodal residuals, *by construction* we must have

$$\sum_{j \in E} \boldsymbol{\phi}_j = \boldsymbol{\phi}^h \quad (8.24)$$

If there exist *distribution matrices* β_i such that $\phi_i = \beta_i \phi^h$, then

$$\sum_{j \in E} \beta_j = I \quad (8.25)$$

3. $\forall i \in \mathcal{T}_h$ assemble the contributions from all $E \in \mathcal{D}_i$ and compute the nodal values of \mathbf{u}^{n+1} by solving the algebraic system

$$\sum_{E \in \mathcal{D}_i} \phi_i = 0 \quad (8.26)$$

As already observed, these schemes are a straightforward matrix extension of the scalar schemes of definition 7.1.3. Most properties extend quite trivially. In particular, with the same analysis of §7.1.2.1 one proves that

Proposition 8.2.1 (Accuracy of space-time matrix \mathcal{RD}). *Space-time matrix \mathcal{RD} schemes are second-order accurate in space and time if*

$$\phi_i = \mathcal{O}(h^3, \Delta t^2)$$

Moreover, for \mathbf{u}^h , \mathcal{F}^h and \mathcal{S}_h given by (3.9), (3.14) and (3.15), one has

$$\phi^h = \mathcal{O}(h^3, \Delta t^2)$$

Hence, \mathcal{LP} schemes defined by $\phi_i = \beta_i \phi^h$ with β_i uniformly bounded with respect to h , Δt and to the data of the problem, are second-order accurate.

While the definition of linear space-time matrix schemes is a trivial extension of definition 8.1.2, the characterization of properties such as energy and L^∞ stability is quite hard. For some schemes, simple results can be obtained by using the available tools. However, the general study of these properties is very difficult and will not be undertaken. Conversely, we are able to extend the definition of a space-time- \mathcal{MU} scheme. The analysis is identical to the one of §7.1.4. Hence, we limit ourselves to report the main differences between the scalar case and the case of the matrix schemes.

We start by observing that if $\mathcal{S} = 0$ the residual (8.23) can be equivalently written as

$$\phi^h = \sum_{j \in E} \left(\frac{\Delta t}{2} K_j + \frac{|E|}{3} I \right) \mathbf{u}_j^{n+1} + \sum_{j \in E} \left(\frac{\Delta t}{2} K_j - \frac{|E|}{3} I \right) \mathbf{u}_j^n = \sum_{j \in E} \left(\tilde{K}_j \mathbf{u}_j^{n+1} + \hat{K}_j \mathbf{u}_j^n \right)$$

where the matrices \tilde{K}_j and \hat{K}_j have been already defined in §3.3 (equation (3.21)), and are a matrix generalization of the Jacobians of the space-time flux of §7.3.1. Making now use of the matrix space-time upwind parameters (3.25), we can define the space-time inflow and outflow states (see equations (7.21) and (7.22))

$$\tilde{\mathbf{u}}_{in} = - \sum_{j \in E} (\tilde{N} \tilde{K}_j^- \mathbf{u}_j^{n+1} + \tilde{N} \hat{K}_j^- \mathbf{u}_j^n) \quad (8.27)$$

and

$$\tilde{\mathbf{u}}_{out} = \sum_{j \in E} (\tilde{N} \tilde{K}_j^+ \mathbf{u}_j^{n+1} + \tilde{N} \hat{K}_j^+ \mathbf{u}_j^n) \quad (8.28)$$

with

$$\tilde{N} = \left(\sum_{j \in E} (\tilde{K}_j^+ + \hat{K}_j^+) \right)^{-1} = - \left(\sum_{j \in E} (\tilde{K}_j^- + \hat{K}_j^-) \right)^{-1} = \frac{1}{2} \left(\sum_{j \in E} (|\tilde{K}_j| + |\hat{K}_j|) \right)^{-1} \quad (8.29)$$

Finally, the space-time residual vector ϕ^h can be written as

$$\phi^h = \left(\sum_{j \in E} (\tilde{K}_j^+ + \hat{K}_j^+) \right) (\mathbf{u}_{out} - \mathbf{u}_{in}) \quad (8.30)$$

We have arrived to an expression formally identical to (7.24), which was used in §7.1.3 to express an analogy with a 1D advection problem along a characteristic line intersecting the prism $E \times [t^n, t^{n+1}]$. As already remarked, for a system the relation between \mathbf{u}_{out} and \mathbf{u}_{in} and the speeds associated to the eigenvalues of the space-time Jacobians is in general very complex and one can hardly see (8.30) as a 1D balance along precise directions in space-time. Nevertheless, we can define schemes which are upwind in the sense that the distribution of the residual to a node i is constrained by the fact that \tilde{K}_j^+ and \hat{K}_j^+ must be nonzero. In particular, the definition of space-time- \mathcal{MU} schemes done in §7.1.4 can be formally extended to the case of the matrix schemes. We do not repeat here the analysis but, with the notation of §7.1.4, we give its final result.

Proposition 8.2.2 (Space-time- \mathcal{MU} matrix \mathcal{RD} schemes and time-marching).

Space-time- \mathcal{MU} matrix \mathcal{RD} schemes are the ones for which on $E \times [t^n, t^{n+1}]$

$$\begin{aligned} \phi_i^{n+1, n+1} &\propto \tilde{K}_i^+ \\ \phi_i^{n, n+1} &\propto \hat{K}_i^+ \end{aligned}$$

Space-time- \mathcal{MU} matrix \mathcal{RD} schemes define a time-marching procedure if

$$\Delta t = t^{n+1} - t^n \leq \min_{E \in \mathcal{T}_h} \min_{j \in E} \frac{2|E|}{3\rho(K_j^+)}, \quad \forall n = 1, M-1 \quad (8.31)$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix.

Since we shall always assume that the *past-shield* condition (8.31) is verified, no confusion is generated by denoting by ϕ_i the local nodal residual for node i . An exception to this rule is given by the two-layer schemes of §7.1.5, for which, however, the matrix extension can be easily obtained making use of formalism introduced here and of the definitions of the next two sections.

8.2.1 Space-time matrix LDA schemes

Matrix variants of the LDA schemes of §7.1.4.1 are obtained by defining distribution matrices formally identical to their scalar counterparts. In particular, in time-dependent computations we refer to the matrix LDA scheme as to the one defined by

$$\phi_i^{\text{LDA}} = \beta_i^{\text{LDA}} \phi^h, \quad \beta_i^{\text{LDA}} = K_i^+ N \quad (8.32)$$

with ϕ^h given by (8.23). Similarly, we refer to the matrix ST-LDA scheme as to the one defined by

$$\phi_i^{\text{LDA}} = \beta_i^{\text{ST-LDA}} \phi^h, \quad \beta_i^{\text{ST-LDA}} = \tilde{K}_i^+ \tilde{N} \quad (8.33)$$

where \tilde{N} is computed as

$$\tilde{N} = \left(\sum_{j \in E} \tilde{K}_j^+ \right)^{-1} \quad (8.34)$$

Note that, in the case of the ST-LDA scheme, this definition of \tilde{N} renders the scheme consistent independently on the past-shield condition. Both the LDA and the ST-LDA schemes are linear and \mathcal{LP} .

8.2.2 Space-time matrix N schemes

In this thesis we consider the matrix formulation of the space-time N schemes of [8, 118] and of [51]. In particular, in unsteady computations, we refer to the matrix N scheme as to the one defined by

$$\begin{aligned} \phi_i^{\text{N}} = \frac{|E|}{3} (\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) &+ \frac{\Delta t}{2} K_i^+ (\mathbf{u}_i^n - \mathbf{u}_{in}^n) \\ &+ \frac{\Delta t}{2} K_i^+ (\mathbf{u}_i^{n+1} - \mathbf{u}_{in}^{n+1}) - \Delta t \beta_i^{\text{LDA}} \sum_{j \in E} \frac{|E|}{3} \mathbf{s}_j \end{aligned} \quad (8.35)$$

with \mathbf{u}_{in} as in (8.9) and β_i^{LDA} as in (8.32). The matrix N scheme is linear and first-order. Being obtained from the N scheme (8.15) with \mathcal{CN} time integration, in the homogeneous case it is unconditionally energy stable (see proposition 8.1.4 and [8, 118]). Similarly, we refer to the ST-N scheme as to the one defined by the local nodal residuals

$$\phi_i^{\text{ST-N}} = \tilde{K}_i^+ (\mathbf{u}_i^{n+1} - \tilde{\mathbf{u}}_{in}) - \beta_i^{\text{ST-LDA}} \sum_{j \in E} \frac{|E|}{3} \mathbf{s}_j \quad (8.36)$$

with $\tilde{\mathbf{u}}_{in}$ as in (8.27) and $\beta_i^{\text{ST-LDA}}$ as in (8.33). Note that, as in the scalar case, if the past-shield condition (8.31) is verified, the ST-N scheme can be written as

$$\phi_i^{\text{ST-N}} = \beta_i^{\text{ST-LDA}} \phi^h + \mathbf{d}_i^{\text{ST-N}}, \quad \mathbf{d}_i^{\text{ST-N}} = \sum_{j \in E} \tilde{K}_i^+ \tilde{N} \tilde{K}_j^+ (\mathbf{u}_i^{n+1} - \mathbf{u}_j^{n+1}) \quad (8.37)$$

The contribution of $\{\mathbf{d}_j^{\text{ST-N}}\}_{j \in E}$ to the energy balance can be shown to be

$$\epsilon^{\text{ST-N}} = \sum_{j \in E} (\mathbf{u}_j^{n+1})^T \mathbf{d}_j^{\text{N}} = \begin{bmatrix} \mathbf{u}_1^{n+1} \\ \mathbf{u}_2^{n+1} \\ \mathbf{u}_3^{n+1} \end{bmatrix}^T D^{\text{ST-N}} \begin{bmatrix} \mathbf{u}_1^{n+1} \\ \mathbf{u}_2^{n+1} \\ \mathbf{u}_3^{n+1} \end{bmatrix}$$

where $D^{\text{ST-N}}$ is the block symmetric matrix

$$D^{\text{ST-N}} = \begin{bmatrix} \tilde{K}_1^+ & 0 & 0 \\ 0 & \tilde{K}_2^+ & 0 \\ 0 & 0 & \tilde{K}_3^+ \end{bmatrix} - \begin{bmatrix} \tilde{K}_1^+ \\ \tilde{K}_2^+ \\ \tilde{K}_3^+ \end{bmatrix} \tilde{N} \begin{bmatrix} \tilde{K}_1^+ \\ \tilde{K}_2^+ \\ \tilde{K}_3^+ \end{bmatrix}^T \quad (8.38)$$

Following [18, 4, 9], one easily shows that $D^{\text{ST-N}}$ is positive semi-definite. Hence, $\{\mathbf{d}_j^{\text{ST-N}}\}_{j \in E}$ define space-time dissipation terms. In particular, the ST-N scheme is more dissipative than the ST-LDA scheme. Starting from the N and ST-N schemes, nonlinear schemes are obtained by formally extending the procedure described in §8.1.3.

8.3 Nonlinear systems

We have finally arrived to our *last stop*: the approximation of weak solutions of

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{u}) = \mathcal{S}(x, y), \quad \text{on } \Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (8.39)$$

on unstructured grids. As for scalar \mathcal{CL} s, when discretizing (8.39) the main problem is the nonlinearity of the relation $\mathcal{F}(\mathbf{u})$, which leads to the formation of discontinuous solutions or to the appearance of *physical* instabilities. The necessity of approximating correctly these features, generates the need of stable and accurate schemes, yielding discrete equations consistent with the mathematical constraints which characterize weak solutions: conservation, in the form of the Rankine-Hugoniot conditions (2.31), and dissipation, in the form of an entropy inequality (2.32). Unfortunately, it is the nonlinearity of the problem which also does not allow a straightforward extension of the schemes developed for scalar advection, as shown by the motivational example of §6.1.

To understand the higher complexity encountered when dealing with a system of nonlinear \mathcal{CL} s we make the following observations. First of all, as anticipated in §3.2.2, we introduce a vector of primary variables $\mathbf{w}(\mathbf{u})$, which in general is different from the set of conserved variables \mathbf{u} . Using this set of variables, we can write (8.39) as the system of quasi-linear PDEs

$$A_0(\mathbf{w}) \frac{\partial \mathbf{w}}{\partial t} + A_1(\mathbf{w}) \frac{\partial \mathbf{w}}{\partial x} + A_2(\mathbf{w}) \frac{\partial \mathbf{w}}{\partial y} = \mathcal{S}(x, y) \quad \text{on } \Omega_T \quad (8.40)$$

where the matrices $\{A_j\}_{j=0,2}$ are given by the Jacobians

$$A_0(\mathbf{w}) = \frac{\partial \mathbf{u}}{\partial \mathbf{w}} \quad \text{and} \quad A_1(\mathbf{w}) = \frac{\partial \mathbf{F}}{\partial \mathbf{w}}, \quad A_2(\mathbf{w}) = \frac{\partial \mathbf{G}}{\partial \mathbf{w}}$$

where we recall that \mathbf{F} and \mathbf{G} are the spatial components of \mathcal{F} (see chapter 2). The set of primary variables \mathbf{w} can be chosen on different grounds. For example, we have seen in §2.4, that nonlinear systems of practical interest are normally equipped with a convex entropy pair $(\mathcal{H}(\mathbf{u}), \mathcal{G}(\mathbf{u}))$ such that weak solutions are characterized by

$$\frac{\partial \mathcal{H}(\mathbf{u})}{\partial t} + \nabla \cdot \mathcal{G}(\mathbf{u}) \leq 0, \quad (8.41)$$

with

$$A_0 = \left(\frac{\partial^2 \mathcal{H}(\mathbf{u})}{\partial \mathbf{u}^2} \right)^{-1}$$

symmetric positive definite. One possible choice for \mathbf{w} is then the set of entropy variables introduced in §2.4:

$$\mathbf{w}(\mathbf{u}) = \mathbf{v}(\mathbf{u}) = \frac{\partial \mathcal{H}(\mathbf{u})}{\partial \mathbf{u}}^T$$

The symmetrization theory for first-order systems of conservation laws (see [78, 121] and also [83, 167, 168] and references therein) ensures that under this change of variables system (8.40) is symmetric. As confirmed by published work on the topic, the use of the entropy variables in the approximation of (8.39) can be beneficial to obtain discretizations with an entropy stable character (see *e.g.* [171, 167, 5, 6, 19, 4, 95, 85] and references therein). However, practical issues often require \mathbf{w} to be a set of variables which best represent the thermodynamics of the problem, as for example when dealing with models of multi-component flows with some form of chemistry, or multi-phase flows with phase transition. There are, in fact, a large number of other possible choices, which depend on the application and on the physics actually modeled by (8.39). In all the cases, the nonlinearity of the flux tensor can pose serious problems in the discretization process. In the case of the matrix \mathcal{RD} schemes considered here, this can be easily shown as follows. Suppose to be seeking a solution of the steady limit of (8.39) in the homogeneous case, by means of the matrix \mathcal{RD} approach of §8.1. As seen in §6.1 and §6.2, the need of correctly reproducing at the discrete level the jump conditions (2.31), imposes that, given the approximation \mathbf{w}_h (equation (3.11)), the residual on an element E must be equal to (see §6.1, §6.2 and definition 6.2.1)

$$\phi^h = \oint_{\partial E} \mathcal{F}(\mathbf{w}_h) \cdot \hat{n} \, dl$$

Since the matrix schemes have been developed for a linear system, we can try to use Gauss's theorem to make the quasi-linear form (8.40) appear in the definition of ϕ^h :

$$\phi^h = \oint_{\partial E} \mathcal{F}(\mathbf{w}_h) \cdot \hat{n} \, dl = \int_E \nabla \cdot \mathcal{F}(\mathbf{w}_h) \, dx \, dy = \int_E \left(A_1(\mathbf{w}_h) \frac{\partial \mathbf{w}_h}{\partial x} + A_2(\mathbf{w}_h) \frac{\partial \mathbf{w}_h}{\partial y} \right) dx \, dy$$

If we introduce the mean-value Jacobians

$$\overline{A}_1 = \frac{1}{|E|} \int_E A_1(\mathbf{w}_h) \, dx \, dy \quad \text{and} \quad \overline{A}_2 = \frac{1}{|E|} \int_E A_2(\mathbf{w}_h) \, dx \, dy \quad (8.42)$$

the linearity of \mathbf{w}_h leads to

$$\phi_h = \sum_{j \in E} \overline{K}_j \mathbf{w}_j$$

where the \overline{K}_j matrices are computed making use of the mean-value Jacobians. Last expression is identical to (8.5), which leads to the conclusion that we can easily apply the matrix \mathcal{RD} schemes of §8.1. However, there is a catch, hidden in the fact that

in general we do not know how to compute the mean-value Jacobians in closed form. The only case in which this is easily accomplished is when the components of \mathcal{F} are quadratic functions of the components of \mathbf{w} . In this case, evaluating the Jacobians in the state given by the arithmetic average of $\{\mathbf{w}_j\}_{j \in E}$ gives exact mean-value flux Jacobians. This fact is used in practically all the \mathcal{RD} literature to solve the Euler equations of a perfect gas (see chapter 10), for which the flux components are quadratic functions of the components of the Roe parameter \mathbf{z} [56, 147]. This means that, in this case, \mathbf{w} must be chosen to be the Roe parameter, being this the only choice leading to conservative schemes. This represents a considerable limitation. In [4], this limitation has been overcome with the introduction of the \mathcal{QRD} formulation of the matrix schemes, in which the exact mean-value Jacobians are replaced by approximate mean-value flux Jacobians, obtained by discretizing (8.42) in NQ Gaussian points. As discussed in §6.2.1, this approach does represent a solution to the problem, but a very expensive one, considering that it involves the evaluation of matrix functions $A_j(\mathbf{w}_h)$ in several points per element. The results presented in [4] show that NQ might have to be quite large to be able to correctly approximate strong discontinuities. For example for a Mach 3.5 blunt body Euler flow, at least 7 quadrature points per element are needed to obtain results comparable to the ones obtained with the use of the Roe parameter. Conversely, the \mathcal{CRD} approach, introduced by [50] and presented in §6.2.2 and §7.2.2, gives a framework which, while being much less mathematically sound, allows to solve the problem in a more efficient way.

In the next section we will recall how the \mathcal{CRD} approach is extended to the matrix schemes of §8.1 and then discuss its extension to the space-time matrix \mathcal{RD} schemes.

8.3.1 \mathcal{CRD} schemes for steady systems of \mathcal{CLs}

The analysis made in the previous chapters, and the matrix approach introduced in §8.1 and §8.2, make easy the task of describing the extension of the \mathcal{CRD} schemes to systems. We consider in this section the case in which steady-state solutions of (8.39) are sought, that is the limit $t_f \rightarrow \infty$. In this case, we proceed exactly as in §8.1, except that the residual (step 1. in §8.1, equations (8.4) and (8.5)) is computed as follows:

$$\phi^h = \sum_{l_j=1}^3 \mathcal{F}^{l_j} \cdot \vec{n}_{l_j} - \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad \text{with} \quad \mathcal{F}^{l_j} = \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}_h(x_p, y_p), \quad (x_p, y_p) \in l_j \quad (8.43)$$

where l_1 , l_2 and l_3 are the edges of E , \vec{n}_{l_j} is the exterior normal to l_j , scaled by the length of the edge, and ω_p is the weight of the p -th quadrature point on l_j . As the one defined by equation (6.6) in §6.2.2, this residual is a *direct approximation* of

$$\phi^h = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n} \, dl - \int_E \mathcal{S}_h \, dx \, dy$$

where the integral of \mathcal{S}_h (equation (3.15)) has been evaluated exactly, while \mathcal{F}_h is a consistent, continuous approximation of the flux. In particular, from now on, we will

assume that $\mathcal{F}_h = \mathcal{F}(\mathbf{w}_h)$, with \mathbf{w}_h as in (3.11). As a consequence, we have that

$$\mathcal{F}_h(x_p, y_p) = \mathcal{F}(\mathbf{w}_h(x_p, y_p))$$

As remarked in §6.2.2, this choice is not necessarily optimal, in the sense that the accuracy and consistency requirements are already satisfied by a piecewise linear approximation of the flux, which can be integrated exactly by using the trapezium rule in (8.43). Anyway, definition (8.43) guarantees that, once a consistent distribution strategy is chosen, a discrete analog of the jump conditions (2.31) is satisfied. Moreover, provided that \mathcal{F}_h is continuous, the schemes respect a Lax-Wendroff theorem ensuring that, if convergent, they converge to the correct weak solutions [6, 5, 9]. We are now only missing a consistent definition of the local nodal residuals ϕ_i . This is the subject of the next sections.

8.3.1.1 The \mathcal{CRD} LDA scheme

As in all the other cases, once the element residual is properly defined, one can immediately obtain a consistent distribution by employing a \mathcal{LP} scheme. In the case of the LDA scheme, this distribution is defined by

$$\phi_i^{\text{LDA}-\mathcal{CRD}} = \beta_i^{\text{LDA}} \phi^h, \quad \beta_i^{\text{LDA}} = K_i^+(\bar{\mathbf{w}})N(\bar{\mathbf{w}}) \quad (8.44)$$

where the definitions of K_i^+ and N are formally identical to (3.16) and (8.11), except that now, due to the nonlinearity of the system, they depend of how the flux Jacobians in (8.40) are linearized. Here we observe that, since by construction we have

$$\sum_{j \in E} \beta_j^{\text{LDA}} = \mathbf{I}$$

and due to the conservative definition of the residual, the issue of discrete conservation is not a problem. Hence, any arbitrary linearization of the flux Jacobian will do. In particular, in (8.44) we have introduced the state $\bar{\mathbf{w}}$, an arbitrary average of \mathbf{w}_h over the element, used to linearize A_1 and A_2 in (8.40). While in §8.1 the linearity of the system has allowed to postulate the existence of N independently on the solution, in the nonlinear case this is not possible anymore. In order for $N(\bar{\mathbf{w}})$ to exist, the matrix

$$N(\bar{\mathbf{w}})^{-1} = \sum_{j \in E} K_j^+(\bar{\mathbf{w}}) \quad (8.45)$$

must be non-singular. This condition has different consequences, depending on the definition of $\mathcal{F}(\mathbf{w}) = \mathcal{F}(\mathbf{w}(\mathbf{u}))$. In particular, for all the systems of equations considered later $N(\bar{\mathbf{w}})^{-1}$ is singular in *stagnation points*, that is, points in which the flow speed vanishes. However, for a symmetrizable system, the well-posedness of the LDA scheme has been proved in [3, 9]. From the practical point of view, this means that in these singular points a *fix* has to be implemented to guarantee that, when inverting numerically the matrix (8.45) and computing β_i^{LDA} , meaningful results are obtained. We recall that the \mathcal{CRD} LDA scheme is \mathcal{LP} , hence second-order accurate.

8.3.1.2 The \mathcal{CRD} N scheme

As in §6.2.2.2, we define the \mathcal{CRD} matrix N scheme by exploiting relation (8.19). In particular, we will refer to the \mathcal{CRD} N scheme as to the one defined by

$$\phi_i^{\text{N-}\mathcal{CRD}} = \phi_i^{\text{LDA-}\mathcal{CRD}} + \mathbf{d}_i^{\text{N-}\mathcal{CRD}} \quad (8.46)$$

with $\phi_i^{\text{LDA-}\mathcal{CRD}}$ as in (8.44), and with

$$\mathbf{d}_i^{\text{N-}\mathcal{CRD}} = \sum_{j \in E} K_i^+(\bar{\mathbf{w}}) N(\bar{\mathbf{w}}) K_j^+(\bar{\mathbf{w}}) (\mathbf{w}_i - \mathbf{w}_j) \quad (8.47)$$

One easily checks that this definition is equivalent to the original one of [50]. Taking $\mathbf{w} = \mathbf{v}$, the entropy variables, the local entropy production of the scheme can be derived with a procedure formally identical to the one used in §6.3 and §6.3.2.2 (see also [4])

$$\Phi_{\mathcal{H}}^{\text{N-}\mathcal{CRD}} = \int_E \bar{\mathbf{v}}_h^T \nabla \cdot \mathcal{F}(\mathbf{v}_h) dx dy + \epsilon^{\text{N-}\mathcal{CRD}} \quad (8.48)$$

where, dropping for clarity the dependence of all the (symmetric) matrices on the linearization, $\bar{\mathbf{v}}_h$ is the piecewise constant function

$$\bar{\mathbf{v}}_h(x, y) = \sum_{E \in \mathcal{T}_h} \sum_{j \in E} \chi_E(x, y) K_j^+ N \mathbf{v}_j$$

and $\epsilon^{\text{N-}\mathcal{CRD}}$ is given by

$$\epsilon^{\text{N-}\mathcal{CRD}} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{bmatrix}^T D^{\text{N-}\mathcal{CRD}} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{bmatrix}$$

with $D^{\text{N-}\mathcal{CRD}}$ the block symmetric matrix

$$D^{\text{N-}\mathcal{CRD}} = \begin{bmatrix} K_1^+ & 0 & 0 \\ 0 & K_2^+ & 0 \\ 0 & 0 & K_3^+ \end{bmatrix} - \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix} N \begin{bmatrix} K_1^+ \\ K_2^+ \\ K_3^+ \end{bmatrix}^T$$

As in the linear case, also in this linearized case one can show that $D^{\text{N-}\mathcal{CRD}}$ is positive semi-definite [18, 4, 9], hence $\epsilon^{\text{N-}\mathcal{CRD}} \geq 0$ represents an entropy dissipative term. As remarked at the end of §6.3.2.2, the behavior of $\bar{\mathbf{v}}_h$ as $h \rightarrow 0$ is not well understood. Hence the balance (8.48) does not represent a stability estimate, even though it implies the existence of an entropy dissipation mechanism. We recall that in [4, 9] it is shown that the matrix N scheme obtained by using the exact mean-value Jacobian linearization is entropy dissipative (see definition 6.3.3 and proposition 6.3.7). While this also does not exactly correspond to entropy stability, it does imply a stronger form of entropy dissipation. We underline that the quantity $\epsilon^{\text{N-}\mathcal{CRD}}$, as well as the dissipation obtained by using the exact mean-value linearization (see equation (6.24) in §6.3.2.2 and [4, 9]), depends on the solution via the linearized flux Jacobians used to evaluate the K_j^+ matrices. Experience has shown that, in some cases, the matrix N

scheme can produce solutions which notably violate the entropy inequality: expansion shocks. This problem has been analyzed in [155] where a cure is also proposed. The situation is analog to what happens in one space dimension with Roe's \mathcal{FV} scheme [147]. Only away from sonic points, this scheme is entropy stable [171], while a fix is needed to carry on this property also in sonic points [171, 125, 82]. The difference with Roe's scheme is that in the case of the N scheme these *entropy unstable* solutions have a truly multidimensional character. However the essence of the phenomenon is the same.

Concerning the L^∞ stability of this scheme, we are unable to present any result. The use of the wave decomposition approach of [10, 9, 118] does not lead anywhere close to a stability (or instability) estimate, even using the equivalence of proposition 6.2.2. This is mainly due to the difference between the exact mean-value Jacobians and the ones used in the distribution. Evidence of the robustness of the \mathcal{CRD} N scheme has been given in [50], and will be confirmed by the numerical results of the next chapters. Lastly, we recall that, for a symmetrizable system, the well-posedness of the N scheme has been proved in [3, 9].

8.3.1.3 Nonlinear schemes

A conservative, nonlinear \mathcal{LP} scheme is obtained by formally extending the procedure of §8.1.3. The scheme obtained by applying this technique to the \mathcal{CRD} N scheme is referred to the \mathcal{CRD} limited N scheme (\mathcal{CRD} LN). Note that, due to the conservative character of the \mathcal{CRD} N scheme, the sufficient well-posedness condition (5.68) is verified. Concerning the entropy stability of the scheme obtained in this way, the remarks of §8.1.3 apply here as well: we do not know how to formally characterize the dissipation properties of the scheme. Conversely, its robustness and ability to preserve the monotonicity of the solution will be proved numerically in the next chapters.

8.3.2 Space-time \mathcal{CRD} schemes

The \mathcal{CRD} formulation of the space-time matrix \mathcal{RD} described in §8.2 has been initially presented in [141, 142]. Time dependent solutions of (8.39) are approximated with the procedure of §8.2, except that the space-time residual ϕ^h is defined by using the \mathcal{CL} form of the problem. In particular, with the notation of (3.13), (3.11) and (8.43), we introduce the quantities

$$\begin{aligned}\mathcal{F}^n &= \sum_{l_j=1}^3 \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(\mathbf{u}^n(x_p, y_p)) \cdot \vec{n}_{lj} - \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \\ \mathcal{F}^{n+1} &= \sum_{l_j=1}^3 \sum_{p=1}^{\text{NC}} \omega_p \mathcal{F}(\mathbf{u}^{n+1}(x_p, y_p)) \cdot \vec{n}_{lj} - \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j\end{aligned}\tag{8.49}$$

We then compute the space-time residual over the prismatic element $E \times [t^n, t^{n+1}]$ as

$$\phi^h = \sum_{j \in E} \frac{|E|}{3} (\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) + \frac{\Delta t}{2} (\mathcal{F}^n + \mathcal{F}^{n+1}) \quad (8.50)$$

It is trivial to verify that ϕ^h is a direct approximation of

$$\phi^h = \int_E (\mathbf{u}^{n+1} - \mathbf{u}^n) dx dy + \int_{t^n}^{t^{n+1}} \oint_{\partial E} \mathcal{F}^h \cdot \hat{n} dl dt$$

with \mathbf{u}^{n+1} , \mathbf{u}^n and \mathcal{F}^h as in (3.13) and (3.14), respectively. This definition alone guarantees that, once a consistent distribution strategy of the residual is chosen, the schemes respect a discrete analog of the jump conditions (2.31). As before, we remark that the approximation (3.14) of the flux can be replaced by a continuous piecewise linear one, without loss of accuracy and still retaining the basic conservative character of the definition. In the following sections we describe the extension of the space-time matrix LDA and N schemes of §8.1.1 and §8.1.2 to this \mathcal{CRD} framework. Nonlinear schemes are again obtained by a formal extension of the procedure described in §8.1.3.

8.3.2.1 Space-time \mathcal{CRD} LDA schemes

Being \mathcal{LP} the space-time LDA and ST-LDA schemes of §8.1.1 immediately extend to this conservative framework. Even though we will not present numerical results with these schemes, this extension is presented here for completeness. In particular, with reference to a generic prismatic space-time element $E \times [t^n, t^{n+1}]$, the \mathcal{CRD} space-time matrix LDA scheme is the one defined by

$$\phi_i^{\text{LDA-}\mathcal{CRD}} = \beta_i^{\text{LDA}} \phi^h, \quad \beta_i^{\text{LDA}} = K_i^+(\bar{\mathbf{u}}) N(\bar{\mathbf{u}}) \quad (8.51)$$

where $\bar{\mathbf{u}}$ is an arbitrary local average of \mathbf{u}^h over $E \times [t^n, t^{n+1}]$. Similarly, the \mathcal{CRD} space-time matrix ST-LDA scheme is defined by the local nodal residual

$$\phi_i^{\text{ST-LDA-}\mathcal{CRD}} = \beta_i^{\text{ST-LDA}} \phi^h, \quad \beta_i^{\text{ST-LDA}} = \tilde{K}_i^+(\bar{\mathbf{u}}) \tilde{N}(\bar{\mathbf{u}}) \quad (8.52)$$

with $\tilde{K}_i^+(\bar{\mathbf{u}})$ is formally given by (3.25), while $\tilde{N}(\bar{\mathbf{u}})$ is computed according to (8.34), which, as already remarked in §8.2.1, renders the ST-LDA consistent, independently on the past-shield condition. Note that, while being conservative due to the definition of ϕ^h , both schemes are \mathcal{LP} and linear. They have no L^∞ stability properties whatsoever. As for the computational example of §7.3.2, we were not able to obtain results with these schemes practically on all the tests discussed in the following chapters.

8.3.2.2 Space-time \mathcal{CRD} N schemes

The space-time matrix N schemes are at the basis of most of the calculations we will present. Their definition makes use of the fact that they can be written as the \mathcal{LP}

LDA scheme plus some dissipation terms. In particular, introducing the quantities

$$\mathbf{d}_i^n = \sum_{j \in E} K_i^+ N K_j^+ (\mathbf{u}_i^n - \mathbf{u}_j^n), \quad \mathbf{d}_i^{n+1} = \sum_{j \in E} K_i^+ N K_j^+ (\mathbf{u}_i^{n+1} - \mathbf{u}_j^{n+1}) \quad (8.53)$$

in time-dependent computations we refer to the \mathcal{CRD} matrix N scheme as to the one defined by

$$\phi_i^{N-\mathcal{CRD}} = \frac{|E|}{3} (\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \frac{\Delta t}{2} (K_i^+ N \mathcal{F}^n + K_i^+ N \mathcal{F}^{n+1}) + \frac{\Delta t}{2} (\mathbf{d}_i^n + \mathbf{d}_i^{n+1}) \quad (8.54)$$

with \mathcal{F}^n and \mathcal{F}^{n+1} given by (8.49). One easily checks that this scheme corresponds to the \mathcal{CRD} N scheme (8.46), combined with trapezium time scheme. As shown by the analysis of §6.3.3.3, which is easily extended to the system case, the entropy dissipation properties of this time integration scheme are not well defined, so that nothing can be said to this regard. However, we know from §8.3.1.2 that the terms \mathbf{d}_i^n and \mathbf{d}_i^{n+1} do introduce entropy dissipation. Note also that in (8.53) and (8.54) the dependence of the matrices on the *arbitrary* average used to linearize the flux Jacobians has been omitted for brevity. By construction, the scheme is consistent, hence conservative, independently on the choice of the average. We will comment on this choice later, when discussing the details relative to the implementation of the schemes.

We also define the following space-time matrix \mathcal{CRD} ST-N scheme:

$$\phi_i^{\text{ST-N-}\mathcal{CRD}} = \tilde{K}_i^+ \tilde{N} \phi^h + \mathbf{d}_i^{\text{ST-N-}\mathcal{CRD}} \quad (8.55)$$

with

$$\mathbf{d}_i^{\text{ST-N-}\mathcal{CRD}} = \sum_{j \in E} \tilde{K}_i^+ \tilde{N} \tilde{K}_j^+ (\mathbf{u}_i^{n+1} - \mathbf{u}_j^{n+1}) \quad (8.56)$$

where, differently from (8.37), all the matrices depend on the linearization used for the flux Jacobians. The analysis of §8.2.2 guarantees that $\mathbf{d}_i^{\text{ST-N-}\mathcal{CRD}}$ has an entropy dissipative character. Also in the case of the \mathcal{CRD} ST-N scheme, the choice of the linearization will be discussed later.

The following remarks are instead very important. First of all, we have no formal evidence of the L^∞ -stable character of these schemes. This will be shown by the numerical experiments of the next chapters. Also, we recall that, for the \mathcal{CRD} N scheme, the need of computing the matrix N might cause problems in stagnation points. We recall that the well-posedness of the scheme has been proved in [3, 9]. This implies that some kind of *fix* has to be implemented in these singular points to obtain meaningful results. Conversely, due to the definition of \tilde{K}_j (equation (3.21)), the parameter matrix \tilde{N} is always well defined.

8.3.2.3 Nonlinear schemes

Conservative, nonlinear \mathcal{LP} schemes are again obtained by formally extending the procedure of §8.1.3. The schemes obtained by applying this technique to the \mathcal{CRD}

N and ST-N schemes are referred to as the \mathcal{CRD} limited N (\mathcal{CRD} LN) and \mathcal{CRD} limited ST-N (\mathcal{CRD} LST-N) scheme respectively. Due to the conservative character of the linear schemes, the sufficient well-posedness condition (5.68) is always verified. The results of the following chapters show that the \mathcal{CRD} LN and LST-N schemes are extremely robust and do give non-oscillatory approximations of time-dependent weak solutions of (8.39).

8.4 Summary

This chapter completes the construction of schemes for (8.39), by extending the \mathcal{CRD} schemes presented in chapters 6 and 7 to the solution of systems of \mathcal{CL} s. The main concepts introduced are summarized hereafter.

- The matrix \mathcal{RD} formulation of [179, 178, 177] for the solution of steady linear hyperbolic systems of PDEs has been recalled. The extension of properties such as linearity, second-order of accuracy and \mathcal{MU} has been described, and the matrix LDA and N schemes have been presented. The matrix N scheme is energy stable in the semi-discrete case, and it is unconditionally energy stable when integrated with the implicit BE scheme and with the \mathcal{CN} scheme;
- We recalled the L^∞ stability criterion of [10, 118, 9]. According to this criterion, the matrix N scheme is L^∞ stable on simple waves. This framework is also used to provide a technique for the construction of limited matrix \mathcal{RD} schemes;
- For time dependent computations, a space-time matrix \mathcal{RD} framework is introduced. The extension of properties such as linearity, second-order of accuracy and space-time- \mathcal{MU} have been recalled. The matrix variants of the space-time LDA, ST-LDA, N and ST-N schemes have been presented.
- The main disadvantage of the matrix \mathcal{RD} formulation is the loss of most of the geometrical analogies which allowed to analyzed the scalar schemes. This is particularly true for \mathcal{MU} and space-time- \mathcal{MU} schemes;
- The extension to nonlinear systems is obtained using the \mathcal{CRD} technique.

Chapter 9

Computational details

In the three following chapters we will show the application of the matrix \mathcal{CRD} schemes proposed in this thesis to the solution of three different systems of \mathcal{CL} s:

Chapter 10: Euler equations of a perfect gas As already observed in §8.3, these equations can be solved by means of a matrix \mathcal{RD} formulation based on a simple conservative mean-value linearization of the flux Jacobians [56], hence the \mathcal{CRD} formulation is not really necessary. However, this chapter allows to test our schemes on problems which are extremely well known, and for which the literature is full of results to which one can refer to for comparison. The content of **Chapter 10** alone constitutes a large database of test-cases and results showing the capabilities of the schemes proposed in the thesis.

Chapter 11: A two-phase flow model We consider a system of \mathcal{CL} s which is a *rough* model of homogeneous two-phase flow. The interest in this system stems from the fact that, while being simple, it is based on thermodynamics which make impossible the use of standard matrix \mathcal{RD} based on a conservative linearization of the Jacobians. This chapter shows one of the many possible fields of applications of our schemes. Our aim is not the development of numerics for multi-phase flow, which is in itself a challenging research field. Thus the tests considered are very academic. However, the results of **Chapter 11** show that the schemes proposed in the thesis can be used in this field.

Chapter 12: Shallow-water equations This is a system of equations of great practical interest. In multiD, no Roe parameter can be found for these equations, which makes the use of our conservative approach well suited. Moreover, this application allows to prove formally and numerically the impressive advantage of the residual approach at the basis of our work. **Chapter 12** represents in itself a very important application of our schemes.

Due to the large variety of applications and to the fact that several different schemes have to be tested, we feel that it is useful to give a summary of the whole discretization procedure, before presenting and discussing the results. In the development of the following chapters, it is hoped that these sections can be used by the reader as a reference to clearly understand what scheme is actually being tested or referred to. Also, this chapter allows to give some more details relative to the implementation of the schemes. We recall that our primary objective is the development of robust residual discretizations for the approximation of weak solutions of systems of conservation laws. Hence, practically all the problems considered involve the formation and interaction of strong discontinuities. For this reason, the next chapters contain no results with the linear \mathcal{LP} LDA schemes. For sake of clarity, the discussion of this chapter is divided into two main parts, summarizing the discretization procedure first for steady calculations and then for time-dependent ones, even though most of the implementation issues are common.

9.1 Steady computations

Hereafter, we summarize in detail the solution procedure used to approximate steady-state solutions of a system of \mathcal{CL} s using the \mathcal{CRD} matrix schemes of §8.3.1. We make use of an explicit iterative scheme that, given the nodal values of an initial solution $\{\mathbf{u}_i^0\}_{i \in \mathcal{T}_h}$, evolves the nodal values as follows

1. $\forall E \in \mathcal{T}_h$ compute the residual using equation (8.43). In all the applications, this formula is intended as an approximation of the contour integral of the discrete flux $\mathcal{F}_h = \mathcal{F}(\mathbf{u}_h^n)$ (see equation (3.11)), with \mathbf{u} used as primary unknown. As a consequence, in the choice of the quadrature formula, we have followed the guidelines emerging in [23] for the evaluation of flux integrals. In particular, a 2-points line Gaussian formula has been used. As remarked in §8.3.1, this is not strictly necessary, since conservation and second of accuracy are already guaranteed by the use of the trapezium rule, equivalent to the exact integration of a linear flux. This probably causes a little overhead in computational time due to the variable interpolation and extra flux evaluations involved. We did not study this aspect, which could be important for the optimization of the schemes. Also we remark that, in our experience, the use of a different set of primary unknowns in the quadrature has little influence on the numerical output;
2. $\forall j \in E$ compute the upwind parameters K_j^+ . To do this, we use the following average state over E :

$$\bar{\mathbf{u}} = \frac{1}{3} \sum_{j \in E} \mathbf{u}_j^n$$

This is the simplest possible choice, from the implementation point of view. As in the case of the flux quadrature, the choice of a different set of variables in the average has little influence on the numerical output;

3. Distribute the residual. We distinguish between the two following possibilities:

CRD N scheme In this case, we compute the local nodal residual ϕ_i as

$$\phi_i = K_i^+(\bar{\mathbf{u}})N(\bar{\mathbf{u}})\phi^h + \sum_{j \in E} K_i^+(\bar{\mathbf{u}})N(\bar{\mathbf{u}})K_j^+(\bar{\mathbf{u}})(\mathbf{u}_i^n - \mathbf{u}_j^n)$$

To allow the computation of $N(\bar{\mathbf{u}})$ also in stagnation points, we add to each diagonal entry of each $K_j^+(\bar{\mathbf{u}})$ the quantity $\epsilon = 10^{-8}$. For each $j \in E$ we also store $\rho(K_j^+)$, the largest eigenvalue of $K_j^+(\bar{\mathbf{u}})$;

CRD limited N scheme In this case we apply the decomposition and limiting procedure described in §8.1.3. The direction $\vec{\xi}$ needed for the decomposition is taken to be the unit vector parallel to the flow speed, if the norm of the latter is larger than 10^{-2} , otherwise we take $\vec{\xi} = (1, 1)/\sqrt{2}$. The limiting is based on the use of mapping (5.65). For each $j \in E$ we also store $\rho(K_j^+)$, the maximum eigenvalue of $K_j^+(\bar{\mathbf{u}})$;

For simplicity, in the following chapters, we will omit the labeling *CRD* and refer to the schemes simply as to the N and limited N (LN) schemes.

4. Boundary conditions. Depending on the problem and on the type of boundary, we might use the following BCs

- Supersonic inlet: In steady computations, for a node i belonging to a supersonic inlet, this normally reduces to setting

$$\sum_{E \in \mathcal{D}_i} \phi_i = 0 \quad (\implies \mathbf{u}_i^{n+1} = \mathbf{u}_i^n)$$

- Symmetry line: Imposed in a strong nodal sense as described in [176];
- Periodicity: Thanks to the fact that for all the grids used in this work the distribution of the nodes along the boundaries is uniform, periodic nodes are treated in strong nodal way, as described in [99];
- Characteristic BCs: Needed for sub-critical inlets and outlets and for inviscid wall BCs. Imposed weakly, as described in [126];

5. $\forall i \in \mathcal{T}_h$ we perform the following explicit update

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\nu}{\lambda_i^+} \sum_{E \in \mathcal{D}_i} \phi_i, \quad \lambda_i^+ = \sum_{E \in \mathcal{D}_i} \rho(K_i^+) \quad (9.1)$$

where the parameter $\nu \leq 1$ is chosen depending on the test;

6. Compute the convergence monitor

$$\|Res(x)\|_{L^1} = \frac{1}{n_{\text{tot}}} \sum_{i \in \mathcal{T}_h} \left| \sum_{E \in \mathcal{D}_i} \phi_i(x) \right|$$

with x a given component of \mathbf{u} , and $\phi_i(x)$ the corresponding component of ϕ_i ;

7. The steps 1.-6. are repeated until $\|Res(x)\|_{L^1}$ is lower than a fixed threshold, or $n + 1$ larger than a fixed limit.

We remark that, while the N scheme always converges to machine accuracy in a limited number of iterations, the LN scheme never reaches this level of convergence. This is a well known problem of all nonlinear *RD* schemes [176, 3, 10].

9.2 Time-dependent computations

In time-dependent computations, starting from the nodal values of the initial solution $\{\mathbf{u}_i^0\}_{i \in \mathcal{T}_h}$, we subdivide the problem of computing the nodal values of the solution at time t_f in a series of explicit iterative loops in space-time slabs $\Omega \times [t^n, t^{n+1}]$. Each loop, known the nodal values $\{\mathbf{u}_i^n\}_{i \in \mathcal{T}_h}$, computes $\{\mathbf{u}_i^{n+1}\}_{i \in \mathcal{T}_h}$ as follows.

1. Computation of the time-step $\Delta t = t^{n+1} - t^n$. this is done in a pre-processing phase. In all the results of the following chapters, Δt has been computed according to the past-shield condition (8.31):

$$\Delta t = 0.75 \min_{E \in \mathcal{T}_h} \min_{j \in E} \frac{2|E|}{3\rho(K_j^+)}$$

where the upwind parameter K_j^+ is computed using the average

$$\bar{\mathbf{u}}^n = \frac{1}{3} \sum_{j \in E} \mathbf{u}_j^n$$

2. Set $\mathbf{u}_i^{n+1,0} = \mathbf{u}_i^n$, $\forall i \in \mathcal{T}_h$. The values $\mathbf{u}_i^{n+1,k+1}$, $k \geq 0$ are computed according to the steps that follow, in which we use the notation $\mathbf{u}_i^{n+1} = \mathbf{u}_i^{n+1,k}$;
3. $\forall E \in \mathcal{T}_h$ compute the space-time residual according to (8.49)-(8.50). As for steady computations, a 2-points line Gaussian formula is used in (8.49), using \mathbf{u} as primary unknown. Also for time-dependent computations we have experimentally seen that, generally, this choice has little influence on the numerical output;
4. $\forall j \in E$ compute the upwind and space-time upwind matrices K_j^+ and \tilde{K}_j^+ . In this case, we use different averages to this purpose. In particular, depending on the distribution scheme selected at run time, we might need to compute one (or some) of the following matrices

$$K_j^+(\bar{\mathbf{u}}^n), K_j^+(\bar{\mathbf{u}}^{n+1}), \tilde{K}_j^+(\bar{\mathbf{u}}^{n+1})$$

where we recall that

$$\bar{\mathbf{u}}^{n+1} = \frac{1}{3} \sum_{j \in E} \mathbf{u}_j^{n+1} = \frac{1}{3} \sum_{j \in E} \mathbf{u}_j^{n+1,k}$$

We add to each diagonal entry of $K_j^+(\bar{\mathbf{u}}^n)$ and $K_j^+(\bar{\mathbf{u}}^{n+1})$ the quantity $\epsilon = 10^{-8}$;

5. Distribute the residual:

CRD N scheme In this case we compute

$$\begin{aligned}\phi_i = & \frac{|E|}{3}(\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \frac{\Delta t}{2} K_i^+(\bar{\mathbf{u}}^n) N(\bar{\mathbf{u}}^n) \mathcal{F}^n + \frac{\Delta t}{2} K_i^+(\bar{\mathbf{u}}^{n+1}) N(\bar{\mathbf{u}}^{n+1}) \mathcal{F}^{n+1} + \\ & \frac{\Delta t}{2} \sum_{j \in E} K_i^+(\bar{\mathbf{u}}^n) N(\bar{\mathbf{u}}^n) K_j^+(\bar{\mathbf{u}}^n) (\mathbf{u}_i^n - \mathbf{u}_j^n) + \\ & \frac{\Delta t}{2} \sum_{j \in E} K_i^+(\bar{\mathbf{u}}^{n+1}) N(\bar{\mathbf{u}}^{n+1}) K_j^+(\bar{\mathbf{u}}^{n+1}) (\mathbf{u}_i^{n+1} - \mathbf{u}_j^{n+1})\end{aligned}$$

with \mathcal{F}^n and \mathcal{F}^{n+1} as in (8.49). For each $j \in E$ we also store $\rho(K_j^+)$, the largest eigenvalue of $K_j^+(\bar{\mathbf{u}}^{n+1})$;

CRD ST-N scheme In this case we compute

$$\phi_i = \tilde{K}_i^+(\bar{\mathbf{u}}^{n+1}) \tilde{N}(\bar{\mathbf{u}}^{n+1}) \phi^h + \sum_{j \in E} \tilde{K}_i^+(\bar{\mathbf{u}}^{n+1}) \tilde{N}(\bar{\mathbf{u}}^{n+1}) \tilde{K}_j^+(\bar{\mathbf{u}}^{n+1}) (\mathbf{u}_i^{n+1} - \mathbf{u}_j^{n+1})$$

For each $j \in E$ we also store $\rho(K_j^+)$;

Limited schemes The limited CRD N scheme and limited CRD ST-N scheme are computed by applying the procedure described in §8.1.3. The direction $\vec{\xi}$ needed for the decomposition is taken to be the unit vector parallel to the flow speed corresponding to \mathbf{u}^{n+1} . If the norm of this speed is less than 10^{-2} , we take $\vec{\xi} = (1, 1)/\sqrt{2}$. The limiting is based on the use of mapping (5.65). For each $j \in E$ we also store $\rho(K_j^+)$;

For simplicity, in the following chapters, we will omit the labeling CRD and refer to the schemes simply as to the N, ST-N, limited N (LN) and limited ST-N (LST-N) schemes.

6. Boundary conditions. The BCs are treated as in steady-state calculations, except that, for some Supersonic inlets, we might impose a known time-dependent nodal value of the solution;
7. $\forall i \in \mathcal{T}_h$ we perform the following explicit update

$$\mathbf{u}_i^{n+1, k+1} = \mathbf{u}_i^{n+1, k} - \frac{0.9}{\lambda_i^+} \sum_{E \in \mathcal{D}_i} \phi_i, \quad \lambda_i^+ = \sum_{E \in \mathcal{D}_i} \left(\frac{|E|}{3} + \frac{\Delta t \rho(K_j^+)}{2} \right) \quad (9.2)$$

8. Compute the convergence monitor

$$\|Res(x)\|_{L^1} = \frac{1}{n_{\text{tot}}} \sum_{i \in \mathcal{T}_h} \left| \sum_{E \in \mathcal{D}_i} \phi_i(x) \right|$$

with x a given component of \mathbf{u} , and $\phi_i(x)$ the corresponding component of ϕ_i ;

9. The steps 3.-8. are repeated until $\|Res(x)\|_{L^1}$ is lower than a fixed threshold, or $k+1$ larger than a fixed limit;
10. If $t^{n+1} = t^n + \Delta t < t_f$ we go back to step 1.

Concerning the iterative convergence, we observe that for the linear schemes a convergence of 3 or 4 orders of magnitude is obtained in few explicit iterations (10 to 20, depending on the problem), while machine accuracy can be obtained in 30 to 40 iterations. With the nonlinear schemes 3 or 4 orders of magnitude of convergence can be obtained in 20 to 30 iterations (depending on the problem). However, machine accuracy is never reached. A similar behavior has been observed in [8, 120] for the limited space-time schemes based on the conservative linearization. We also mention that, in the framework of the space-time schemes of [47, 53], implicit iterative techniques have been experimented in [65]. Generally, implicit solution strategies allow to reach the same fixed level of convergence in less iterations. However, they are also more expensive in terms of CPU time, due to the need of evaluating the Jacobian of the nodal residuals. The general conclusion is that for the type of time-dependent problems considered in next chapters, involving the propagation of discontinuities, an explicit strategy results in a faster convergence in terms of CPU time [65].

9.2.1 Two-layer schemes

We will not describe in detail the solution procedure corresponding to the two-layer version of the schemes (see §7.1.5), used in the few results shown in **Chapter 10**. The matrix \mathcal{CRD} formulation of these schemes can be easily obtained by combining the definitions of §8.3.2.2 and §8.3.2.3 with the ones given in §7.1.5. The interested reader can also refer to [51, 8]. We limit ourselves to observe that, when using this two-layer formulation, the time-step in the second layer can be computed according to two different approaches. One approach is to fix the ratio

$$Q = \frac{\Delta t_2}{\Delta t_1} = \frac{t^{n+2} - t^{n+1}}{t^{n+1} - t^n} \quad (9.3)$$

with Δt_1 computed as in step 1. of §9.2. This leads to

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

Alternatively, one can fix the value of the total time-step $\Delta t = \Delta t_1 + \Delta t_2$, giving

$$Q = \frac{\Delta t}{\Delta t_1} - 1, \quad \Delta t_2 = \Delta t - \Delta t_1$$

with Δt_1 computed as in step 1. of §9.2. In **Chapter 10** we will present results obtained using both the approaches.

Chapter 10

Evaluation on the Euler equations of a perfect gas

We consider here the Euler equations of a perfect gas, which constitute a homogeneous system of \mathcal{CL} s, with conserved variables and flux given by

$$\mathbf{u} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix}, \quad \mathcal{F}(\mathbf{u}) = \begin{bmatrix} \rho u & \rho v \\ \rho u^2 + p & \rho uv \\ \rho uv & \rho v^2 + p \\ \rho Hu & \rho Hv \end{bmatrix} \quad (10.1)$$

where ρ is the fluid density, $\vec{u} = (u, v)$ is the flow speed, E is the total energy per unit mass, and H is the enthalpy per unit mass

$$H = E + \frac{p}{\rho}$$

The system is closed by the perfect gas Equation Of State (EOS)

$$p = (\gamma - 1)\rho \left(E - \frac{\vec{u} \cdot \vec{u}}{2} \right) \quad (10.2)$$

where γ is the ratio of the specific heat coefficients, here assumed to be $\gamma = 1.4$. One easily checks that the components of \mathcal{F} are quadratic functions of the components of the Roe parameter [147]

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

Choosing as primary variables $\mathbf{w}_h = \mathbf{z}_h$, one can construct matrix \mathcal{RD} schemes based on an exact mean-value linearization of the Jacobians of $\mathcal{F}(\mathbf{z})$, obtained in each element E by evaluating these matrices in the arithmetic average of the nodal values

$\{\mathbf{z}_j\}_{j \in E}$. This multidimensional conservative linearization, firstly proposed in [56], is at the basis of most or all of the steady-state computations of flows of a perfect gas in the \mathcal{RD} literature. It has also been used in [8, 118, 120] and [47, 53, 44, 51] to perform time-dependent computations using space-time matrix \mathcal{RD} schemes. Hence, the conservative approach proposed in this thesis is not strictly necessary to solve these equations. However, as shown also in [50], it gives a more *flexible* formulation of the schemes, while guaranteeing a correct approximation of weak solutions. The amount of published literature proposing test-cases involving the solution of the Euler equations is impressive. It is mainly for this reason that we discuss here the results obtained with our schemes on a large number of well established problems, for which reference results are available.

10.1 Steady computations

10.1.1 Mach 10 blunt body flow

In [50] the \mathcal{CRD} matrix schemes have been already tested on quite a number of problems, showing that their performances, when solving the Euler equations, are comparable to the ones of the matrix \mathcal{RD} schemes of [56, 179, 177], based on the conservative Roe linearization. Here, we want to add to the results of [50] one test proving the robustness of the limited N scheme (see §9.1). In particular, we consider a Mach 10 flow about a circular cylinder. A sketch of the problem and of the spatial domain used in the computation is reported on the left on figure 10.1.

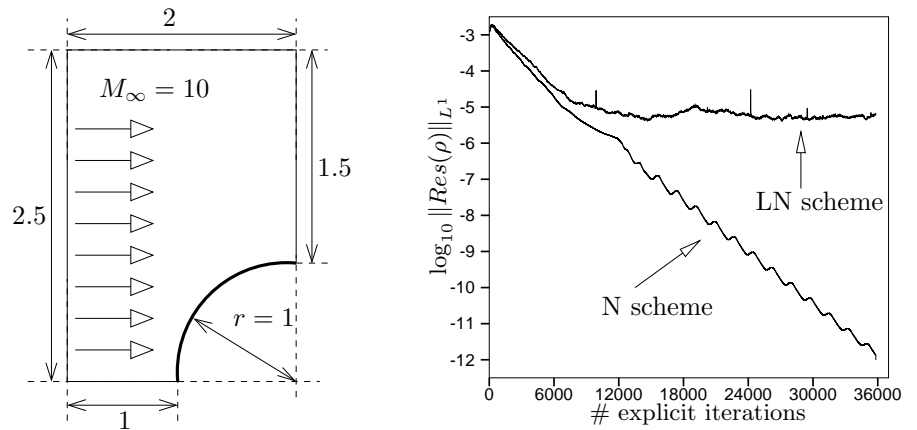


Figure 10.1: Mach 10 bow shock around a circular cylinder. Geometry (left) and convergence histories (right)

Due to the symmetry of the problem, we have simulated only half of the flow, by setting symmetry BCs on the lower boundary. The cylinder is treated as an inviscid wall, while the left boundary is a supersonic inlet. The computation is started from

a uniform Mach 10 flow everywhere, except on the cylinder where the speed normal to the wall has been set to zero. The irregular mesh used for this computation has a reference element size $h = 1/50$. On the right in figure 10.1, we report the explicit convergence histories of the N and LN schemes, in terms of the residual of the density. The iterations are performed with $\nu = 0.5$ (see equation (9.1)). The trends observed in the picture are standard in \mathcal{RD} : the N scheme converges to machine accuracy in a relatively small number of iterations, while the convergence of the limited scheme stalls after a drop of two orders of magnitude. Depending on the test-case, one might be able to obtain a residual drop of three or four orders of magnitude with the nonlinear scheme, but the general behavior is well represented by the plot of figure 10.1. This is a common feature of all nonlinear matrix \mathcal{RD} schemes [176, 3, 10]. We remark that this stall in convergence is not seen in the scalar case, as shown by the results of §5.6.1, §5.6.2 and §6.4. As argued in §9.1 and §9.2, this behavior could be related to a weak energy (entropy) stability. The analysis of §5.5.2.3 shows one possible mechanism that could lead to this effect. For steady scalar problems, however, the beneficial effects of the \mathcal{MU} , especially in 1-target elements (see *e.g.* proposition 6.3.5), seem to be enough to completely stabilize the scheme. For a system, the coupling of the equations introduced by the matrix approach weakens the effects of the \mathcal{MU} , while probably increasing the number of situations in which 2-target or even 3-target elements are reverted to 1-target by the limiting procedure, thus making the effects of the instability shown in §5.5.2.3 more pronounced. What is more surprising is that after the stall, the solution in correspondence of the shock does not change sensibly, most of the *perturbations* being placed in the smooth regions of the flow. The capturing of discontinuities being very stable, the origin of the problem must be sought somewhere else. This aspect is not well understood and it definitely needs to be object of future study.

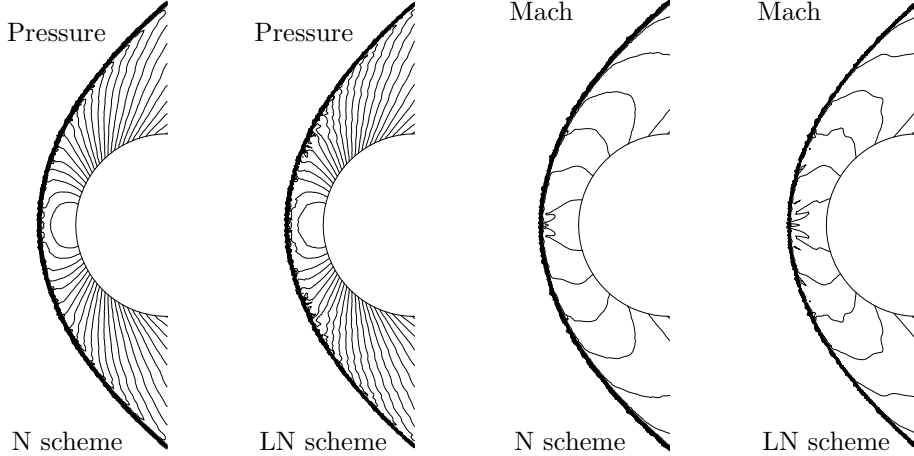


Figure 10.2: Mach 10 bow shock around a circular cylinder. Pressure (left) and Mach (right) contours obtained with the \mathcal{CRD} N scheme and with the \mathcal{CRD} LN scheme

We report on figure 10.2 contour plots of the pressure and Mach number computed with the N and LN schemes. The results of the limited schemes are obtained stopping the computation after 36000 iterations. The contours clearly indicate a stable and quite

monotonic capturing of the bow shock. We recall that in [50] the same computations could not be run with the blended scheme proposed by Deconinck and collaborators (see §5.5.1 and [55, 154, 155, 57]). Similar conclusions are reported in [10, 9, 118] for a Mach 8 bow shock. In the reference, the limited N scheme gives a stable solution while the blended scheme, in this case the one proposed in [3], is not able to handle the strong shock. This observation suggests that the limiting technique really is effective to design discretization with stable discontinuity capturing properties. Moreover, the contours of the solution of the LN scheme show a sharper numerical shock. Qualitatively, our result is comparable to the one reported in [10, 9, 118].

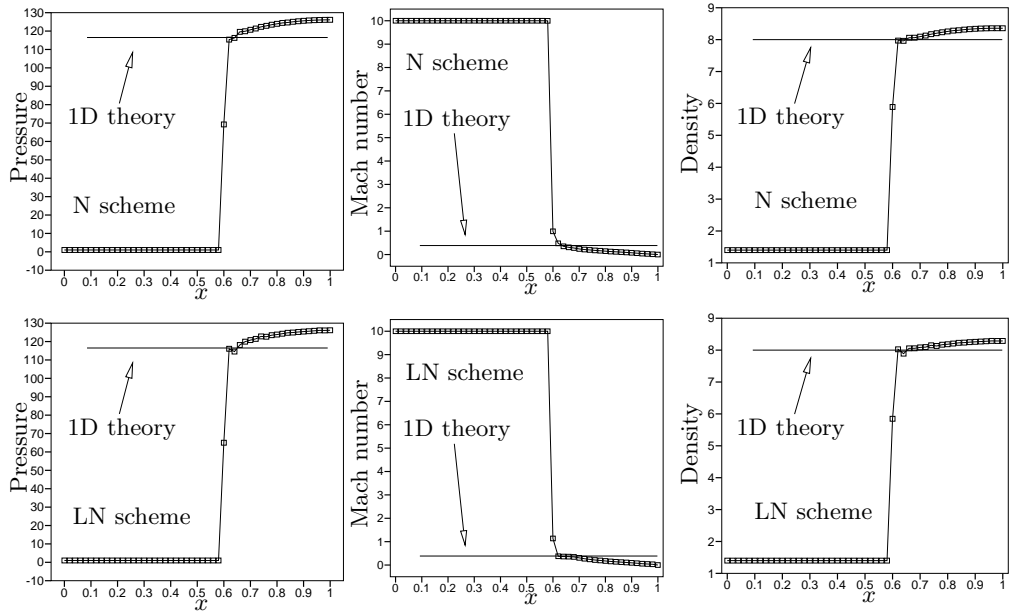


Figure 10.3: Mach 10 bow shock around a circular cylinder. Pressure (left), Mach number (middle) and density (right) distribution along the stagnation line computed with the N scheme (top) and with the LN scheme (bottom). The solid line represents the theoretical post-shock values.

To complete the analysis of the solution, we report on figure 10.3 the data extracted on the stagnation line $y = 0$. In all the plots we also draw the line corresponding to the post-shock values of the variables, computed using the onedimensional jump conditions. Generally, we see that the strength of the numerical shock is correct, confirming the conservative character of the discretization. We can also observe that the N scheme really yields a non-oscillatory solution, while for the LN scheme a small oscillation appears right after the discontinuity, as it can be seen from the pressure and density plots. Given the strength of the shock, these results are judged to be extremely good.

10.2 Time-dependent computations

The most important contribution of this thesis is the extension of the conservative framework of [50] to the time-dependent case. It is for this case that most of the numerical experiments have been carried on.

10.2.1 Mach 10 moving shock

This test is meant to be a time-dependent variant of the computation of §10.1.1. We solve the Euler equations on the spatial domain $\Omega = [0, 2] \times [0, 0.1]$. At time $t = 0$ we prescribe a solution corresponding to a right-moving Mach 10 shock situated at $x = 0.5$. Pressure and density in the initial solution are chosen such that the speed of the shock is equal to 10. We let the schemes compute the movement of the shock until time $t_f = 0.1$, corresponding to a displacement of the *exact* shock of a unit length. The problem has been solved using both an irregular and a structured triangulation (see figure 3.1) with reference element sizes $h = 1/100$. Periodic BCs have been imposed on the top and bottom boundaries, so to obtain a quasi-1D solution. No visible differences are present between the results obtained on the irregular and on the structured grid. In figures 10.4 and 10.5 we visualize the ones obtained on the unstructured mesh with the space-time N, ST-N, LN and LST-N schemes (see §9.2). The pictures show plots of the data extracted along the line $y = 0.05$.

Firstly, we observe that the position of the exact shock is well approximated by the linear N and ST-N schemes and perfectly reproduced by their limited variants. This confirms the conservative character of the schemes. From the plots on figure 10.4, we can also see that the shock layer of the ST-N schemes is wider than the one of the N scheme. This is in agreement with the scalar results of §7.3.1 and §7.3.2: the ST-N scheme has a more diffusive behavior. On the same figure, and on the plots on figure 10.5, we also see that in correspondence of the shock there are no oscillations, and that the limited schemes do give a sharp capturing of the discontinuity. However, all the plots also show the presence of perturbations moving upstream of the shock. We can see two of them in the density profiles, and only one of them in the pressure distributions. The only exception to this is the ST-N scheme for which the two perturbations in the density profile are blurred into a unique feature (see picture on the bottom-right on figure 10.4), due to the diffusive character of the scheme.

These perturbations are clearly due to a discretization error at $t = 0$. To confirm this hypothesis we have run the simulations on finer meshes. Qualitatively, the results of this analysis are summarized on figure 10.6. The left picture shows the results obtained on structured triangulations with the linear N scheme. We plot the nodal values of the density in the middle of the domain, for mesh sizes $h = 1/100$ and $h = 1/500$. As the mesh is refined, a reduction of the error is indeed observed. However, due to the fact that the error is generated at $t = 0$ in correspondence of the initial singularity, this reduction is considerably less than of $\mathcal{O}(h)$. We clearly see that the scheme splits the error in components moving along the characteristic of the Euler equations. In fact,

one easily checks that the perturbation closer to the origin of the x -axis moves with speed $u_L - a_L$, being a_L the speed of sound after the shock. We conclude that this must be the projection of the error generated at $t = 0$ on the characteristic corresponding to the *slow acoustic* speed. Similarly, the second perturbation moves with speed u_L . Hence it is the projection of the error on the entropy field, as confirmed by its absence in the pressure distributions. Probably a third component of the error, which we have been unable to detect, is also present.

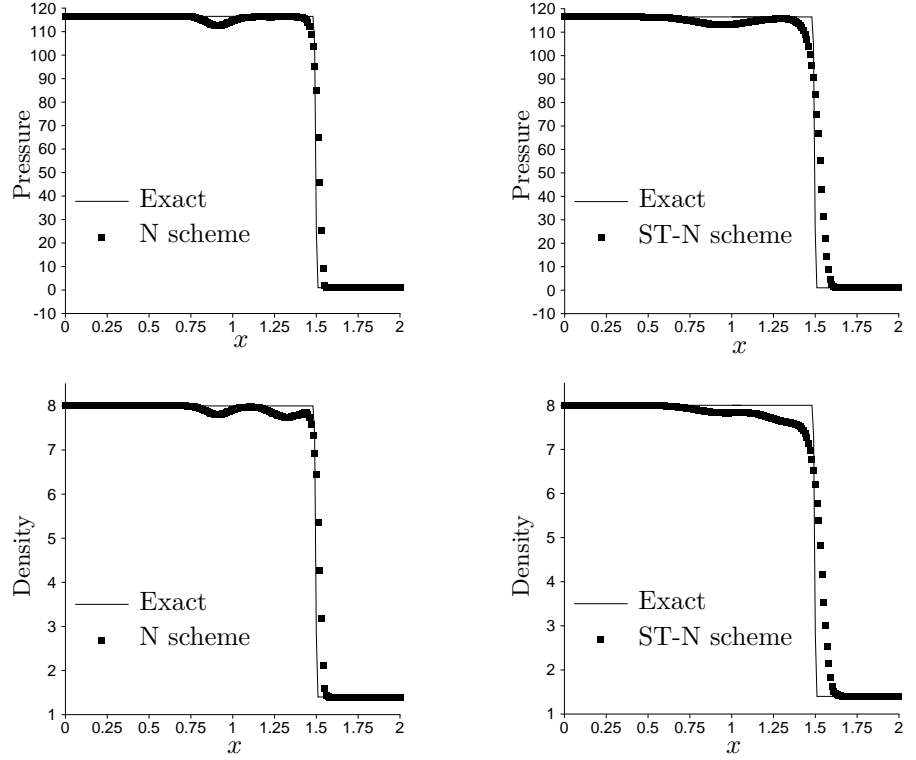


Figure 10.4: Mach 10 shock. Pressure (top) and density (bottom) profiles at time $t = 0.1$. Data extracted at $y = 0.05$. N scheme (left) and ST-N scheme (right)

Similar conclusions can be drawn from the right picture in figure 10.6 where the results obtained with the LN scheme on a structured triangulation with $h = 1/500$ are reported. Comparing this result with the one on figure 10.5, we see again that the reduction of the error is indeed small¹. The analysis has led to similar results for the ST-N and LST-N schemes. These effects, however, are not induced by the conservative approach proposed here. We performed the same computations using the space-time N scheme of [8, 118], based on the conservative Roe linearization, obtaining results identical to the ones on the left in figure 10.4, given by the \mathcal{CRD} N scheme. A similar

¹The error on the finer mesh is about half of the error on the coarse one

behavior is observed also for the space-time schemes of [46, 53, 47]. We attribute the appearance of this error to the difference between the exact jump relations and their piecewise linear approximation on the mesh actually used by the schemes. This difference produces the small amplitude perturbations seen in the results when applying the schemes to the initial *exact* shock. Other explanations could be possible [15, 101] and this behavior certainly deserves a more detailed investigation.

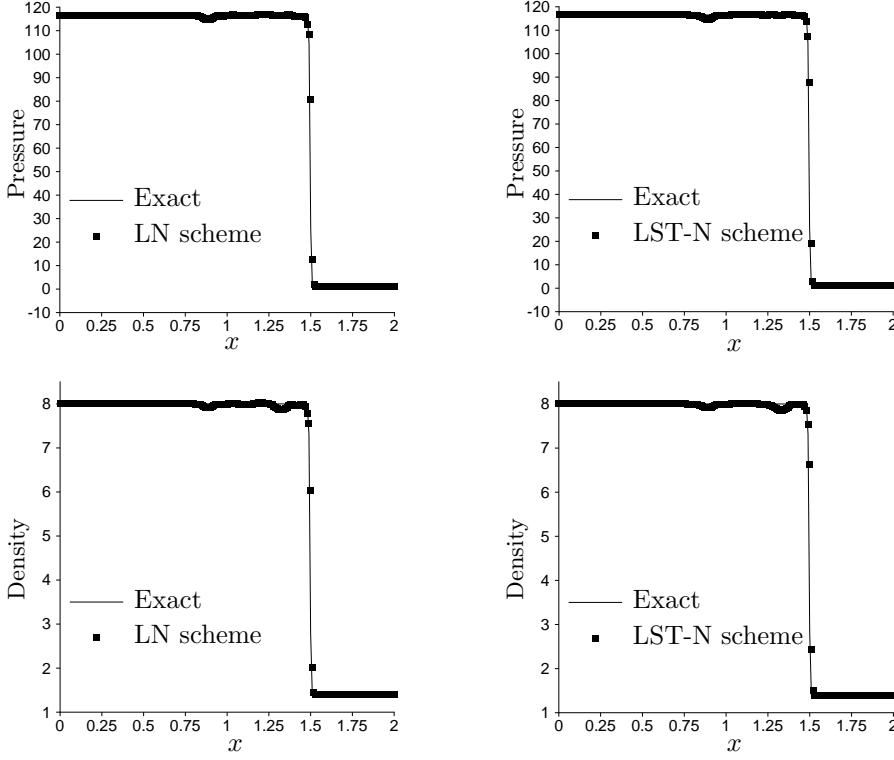


Figure 10.5: Mach 10 shock. Pressure (top) and density (bottom) profiles at time $t = 0.1$. Data extracted at $y = 0.05$. LN scheme (left) and LST-N scheme (right)

We recall that these schemes are based on ideas which are multidimensional variants of the upwinding incorporated by the first-order Roe scheme [147]. It would not be surprising to discover that, in a multidimensional setting, they show flaws similar to the ones suffered by Roe's scheme [135]. Unfortunately, their algebraic complexity is such that an analytical investigation is prohibitively complex. At the moment, our experience indicates a very robust behavior. The results of this section confirm this robustness. In particular, we underline once more the monotone approximation of the discontinuity given by the nonlinear LN and LST-N schemes. This can be clearly seen in the plots on figure 10.5 and in the plot of the *fine mesh* solution on the right in figure 10.6. Considering the strength of the shock simulated, these are very good results.

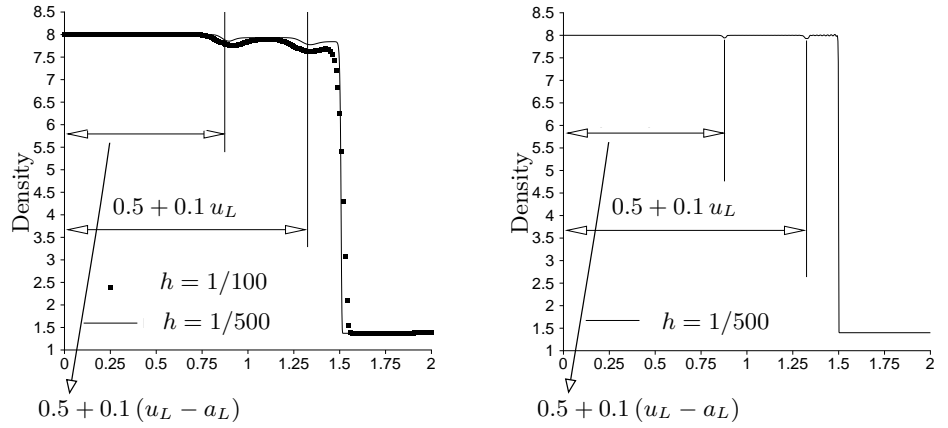


Figure 10.6: Mach 10 shock: error propagation. N (left) and LN (right) scheme

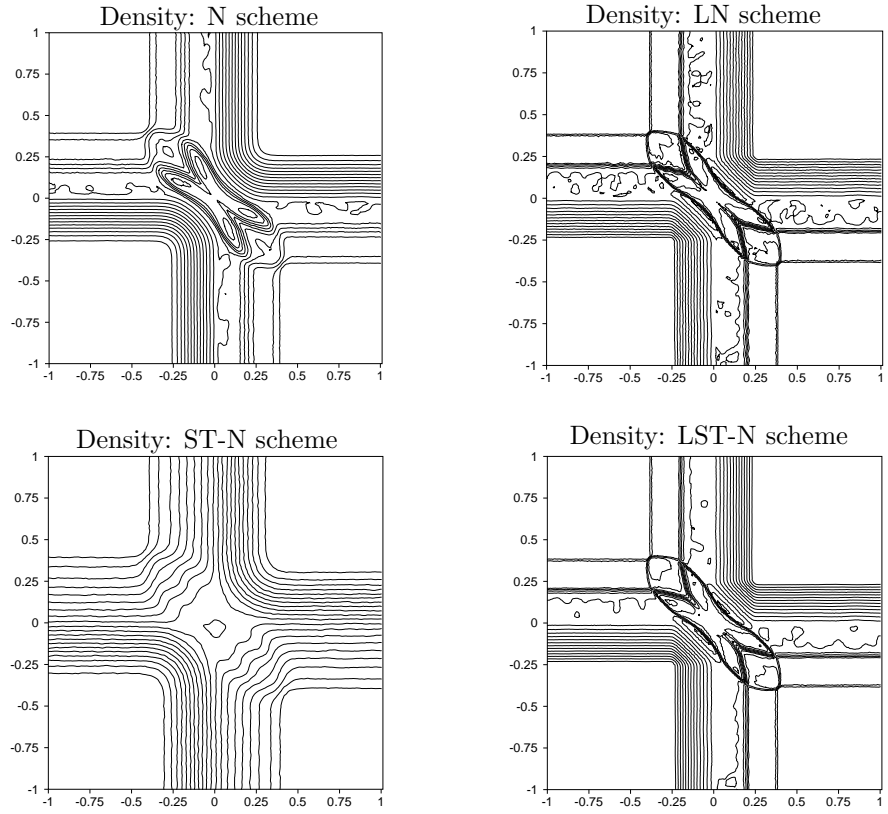


Figure 10.7: 2D Riemann Problem. Density contours at time $t = 0.2$. Top: N (left) and LN scheme (right) - Bottom: ST-N (left) and LST-N scheme (right)

10.2.2 A 2D Riemann problem

This problem is taken from [8]. At time $t = 0$ the velocity is set to zero, and the following discontinuity in pressure and density is imposed:

$$p = \begin{cases} 1 & \text{if } xy \geq 0 \\ 0.1 & \text{otherwise} \end{cases}, \quad \rho = \begin{cases} 1 & \text{if } xy \geq 0 \\ 0.1 & \text{otherwise} \end{cases}$$

We compute the solution up to time $t = 0.2$ with the N, ST-N, LN and LST-N schemes, on an unstructured discretization of the spatial domain $[-1, 1]^2$ with $h = 1/100$ as in [8]. Symmetry BCs are imposed on all the boundaries. Contour plots of the computed density field are given in figure 10.7, while a comparison of the the numerical solutions on the lower boundary of the domain with the exact onedimensional solution of the problem is reported in figure 10.8.

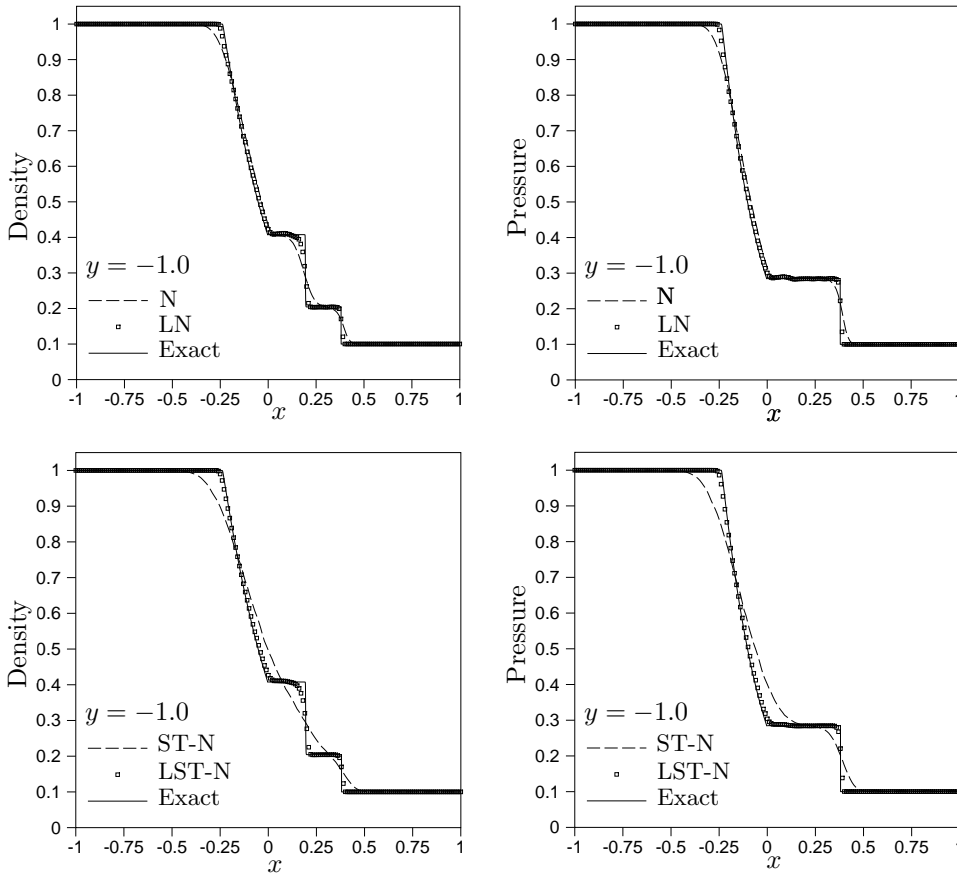


Figure 10.8: 2D Riemann Problem. Density (left) and pressure (right) at $t = 0.2$ and $y = -1.0$. Top: N and LN schemes - Bottom: ST-N and LST-N schemes

All the discontinuities are computed monotonically, they have the proper strength, and are in the correct positions as it can be clearly seen in figure 10.8. From the figures, one notes a striking difference between the results given by the linear ST-N scheme and the nonlinear LST-N one. Similarly, a remarkable difference is observed between the solutions of the linear N and ST-N schemes. Clearly, the ST-N scheme yields the worst results in terms of accuracy. The plots on the bottom on figure 10.8 show that this scheme blurs the three waves almost into a unique smooth curve. These results indicate that the space-time character of the upwinding that this scheme incorporates leads to an excess of numerical diffusion. What is surprising is that, despite of the impressive difference between the N and ST-N schemes, the results of the LN and LST-N schemes are nearly identical. In particular, both produce a very crisp resolution of the wave interactions and a non-oscillatory approximation of the discontinuities. This is very interesting since it suggests that, in the computation of flows containing mainly strong discontinuities, the results given by a limited scheme are qualitatively almost independent on the nature of the underlying linear scheme. This shows the effectiveness of this technology to construct truly shock capturing discretizations. The overall quality of our results is comparable to the one achieved in [8].

10.2.3 Double Mach reflection

This problem, initially proposed in [187], constitutes a severe test for the robustness and the accuracy of schemes designed to compute discontinuous flows containing complex structures. It consists of the interaction of a planar right-moving Mach 10 shock with a 30° ramp. We use a frame of reference in which the x axis is aligned with the ramp. The computational domain is then the square $[0, 3] \times [0, 0.8]$, with the ramp starting at $x = 1/6$. In the initial solution, the shock forms a 60° angle with the x -axis (see figure 10.9). The simulation has been run on an unstructured triangulation with $h = 1/100$ until time $t_f = 0.2$. The exact motion of the shock is imposed on the top boundary.

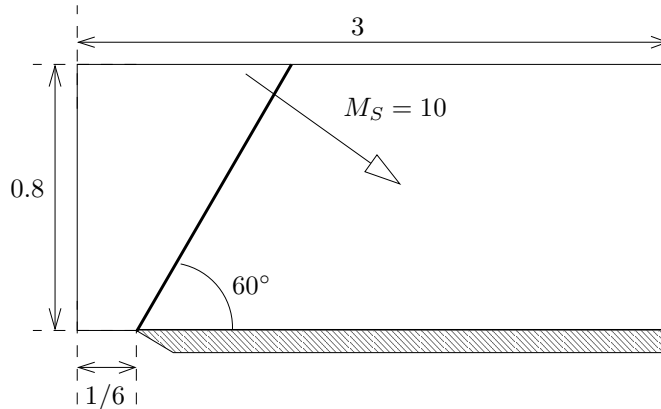


Figure 10.9: Double Mach reflection: sketch of the initial solution

As it is customary for this test, we plot the contours of the density field. In figure 10.10 three results are shown. On the top we report the solution obtained with a second-order cell-centered \mathcal{FV} scheme using Roe's numerical flux, linear reconstruction and limiter proposed in [24], and a second-order Runge-Kutta time integrator. On the middle and bottom pictures, we show the results obtained (on the same mesh) with the nonlinear LN and LST-N schemes respectively. All the schemes resolve quite well the interaction between the shock and the ramp. However, the resolution of the contact emanating from the triple point and of the jet of fluid on the wall improves going from the top to the bottom picture, the \mathcal{FV} scheme giving the worst result. The two limited \mathcal{RD} schemes show a sharper capturing of these features and of the shock. For this test, the nonlinear LN scheme gives the best result. However, to be completely fair to the \mathcal{FV} scheme, we have to mention that for this type of problems the \mathcal{RD} computations take considerably more CPU time, due to their implicit character.

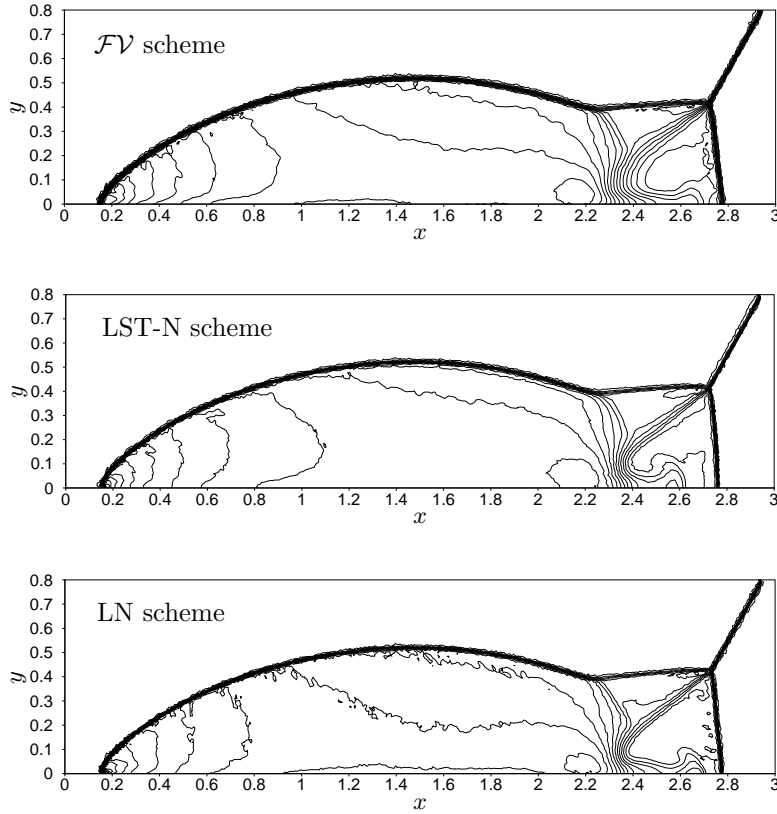


Figure 10.10: Double Mach reflection. Density at time $t = 0.2$. Cell-centered \mathcal{FV} scheme (top), LST-N scheme (middle) and LN scheme (bottom)

10.2.3.1 Grid refinement with the LN scheme

Strong slip lines in inviscid flows can lead to the formation of a Kelvin-Helmholtz (KH) instability. Accurate numerical methods are able to reproduce the formation of such an instability without dissipating it. This has motivated several authors (see *e.g.* [42, 158]) to perform grid refinement studies on problems containing strong contact discontinuities, such as the double Mach reflection. We have conducted such a study with the LN scheme, which is the most performing among the nonlinear \mathcal{RD} schemes we propose. The results are visualized in figure 10.11, where the contours of the density in the vicinity of the triple point are plotted.

On the top pictures in the figure, we report three \mathcal{RD} solutions, obtained on irregular meshes with $h = 1/100$ (top left), $h = 1/200$ (top right) and $h = 1/400$ (second row). As the mesh is refined, the LN scheme remains stable across shocks. The solutions on the finer meshes are quite clean, despite of the irregularity of the grids. Also, the solution obtained for $h = 1/200$ already gives a *glimpse* of the formation of the KH instability. This instability is instead clearly visible on the solution obtained on the finest mesh, even though it has not broken into its typical roll-ups, yet.

On the last row on figure 10.11, we report, for comparison, a solution obtained with a second-order $Q1 - \mathcal{DG}$ scheme (left) on a structured mesh composed of quadrilaterals ($\Delta x = \Delta y = 1/480$), and the solution obtained with a third-order WENO scheme on a triangulation with $h = 1/400$ (right). The \mathcal{DG} and WENO results are taken from [42] and [158], respectively. Comparing the results on the finest meshes, we see that of all the methods the WENO scheme, while being third-order accurate, gives the poorest resolution of the interaction. This proves that schemes having a residual character introduce a markedly lower error. This even with respect to WENO discretizations having a higher accuracy, hence a faster reduction of the error with the mesh size. Note also that the LN scheme gives already a solution comparable to the WENO scheme on the medium mesh ($h = 1/200$). Conversely, the \mathcal{DG} and the \mathcal{RD} solutions have a similar quality. This taking into account that the \mathcal{DG} result is obtained on a structured mesh with considerably smaller element sizes¹.

These results indicate that the \mathcal{RD} technology developed here, and by other authors [8, 10], has the necessary accuracy and robustness to compete with the \mathcal{DG} schemes. However, while our results show a good potential, the design of higher-order discretizations, which is relatively natural in the \mathcal{DG} framework [42], is still an open issue. Even though encouraging preliminary results do exist [12, 9, 139], they show the necessity of a better understanding of the nonlinear mapping technique, used here and in the references to construct non-oscillatory high-order schemes. Moreover, very high-order \mathcal{DG} schemes benefit, in terms of efficiency, from the possibility of using a Lax-Friederichs-type dissipation as main building block of their stabilization. Further understanding of the limiting technique could allow this also for \mathcal{RD} schemes (see §6.2.2.3 and

¹the reduction of the mesh size from $1/400$ to $1/480$ implies an increase of the number of nodes roughly of the 44%!

[10, 9, 118]). We believe that a successful application of this technique will be the key to the success of \mathcal{RD} . The limited schemes, in fact, have a true residual character, while being built on ideas allowing the preservation of the local monotonicity of the solution. The understanding of this procedure will be fundamental for the construction of very high-order \mathcal{RD} discretizations, as well as for the design of more efficient ones.

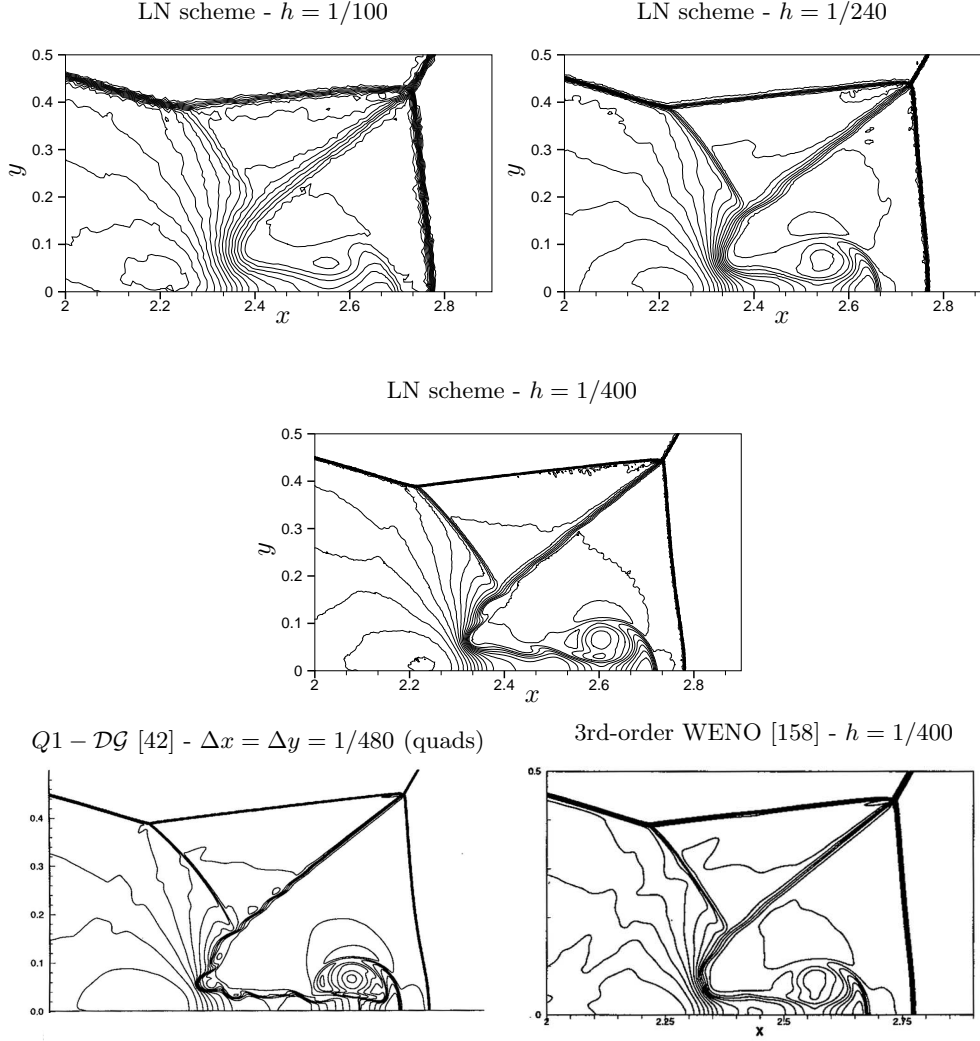


Figure 10.11: Double Mach Reflection. Grid refinement study for the LN scheme. Top row: LN scheme solution for $h = 1/100$ (left) and $h = 1/240$ (right). Middle: LN scheme solution for $h = 1/400$. Bottom row: $Q1 - \mathcal{DG}$ solution (left) on grid of quads with $\Delta x = \Delta y = 1/480$ [42] and third-order WENO scheme (right) on triangulation with $h = 1/400$ [158]

10.2.4 A shock-shock interaction

This test has been included to further assess the shock-capturing capabilities of the nonlinear space-time schemes in a true multidimensional situation. It is one of the two-dimensional Riemann problems studied in [106] and later used also in [111, 60, 61, 62]. The problem consists of the interaction of two oblique shocks with two normal shocks. All the discontinuities are moving backwards with respect to the speed in the pre-shock region as depicted in the sketch on figure 10.12. The spatial domain is the square $[0, 1]^2$.

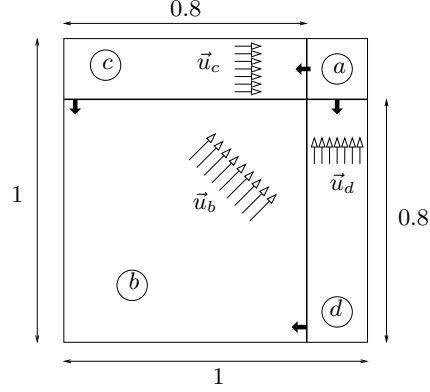


Figure 10.12: Shock-shock interaction. Initial solution

With reference to the notation of figure 10.12, the initial data are given by

$$(\rho, u, v, p) = \begin{cases} (1.5, 0, 0, 1.5) & \text{state } a \\ (0.1379928, 1.2060454, 1.2060454, 0.0290323) & \text{state } b \\ (0.5322581, 1.2060454, 0, 0.3) & \text{state } c \\ (0.5322581, 0, 1.2060454, 0.3) & \text{state } d \end{cases}$$

Due to the symmetry of the problem, only half of the flow has been simulated. In particular, the diagonal of $[0, 1]^2$ has been used as a symmetry line. The computations have been run on an unstructured discretization with reference mesh size $h = 1/200$. Symmetry BCs are imposed also on the top boundary, where a normal moving shock is present, while the movement of the oblique shock on the left boundary has been imposed exactly. We compare the numerical solution obtained with the LN and LST-N schemes with the one of the \mathcal{FV} scheme used in the double Mach reflection test. As in [111, 60, 61] we compute the interaction up to time $t_f = 0.8$ and visualize the results in terms of contours of the density. On figure 10.13, in particular, we visualize from left to right the solutions of the LST-N scheme, of the \mathcal{FV} scheme, and of the LN scheme. The nature of the flow is quite complex. The interaction of the shocks generates two symmetric lambda-shaped couples of shocks and a downward moving normal shock. Very strong slip lines emanate from the lower triple points and interact with one of the branches of the upper lambda-shocks, while a jet of fluid is pushed from the high pressure region (state a in figure 10.12) against the normal shock.

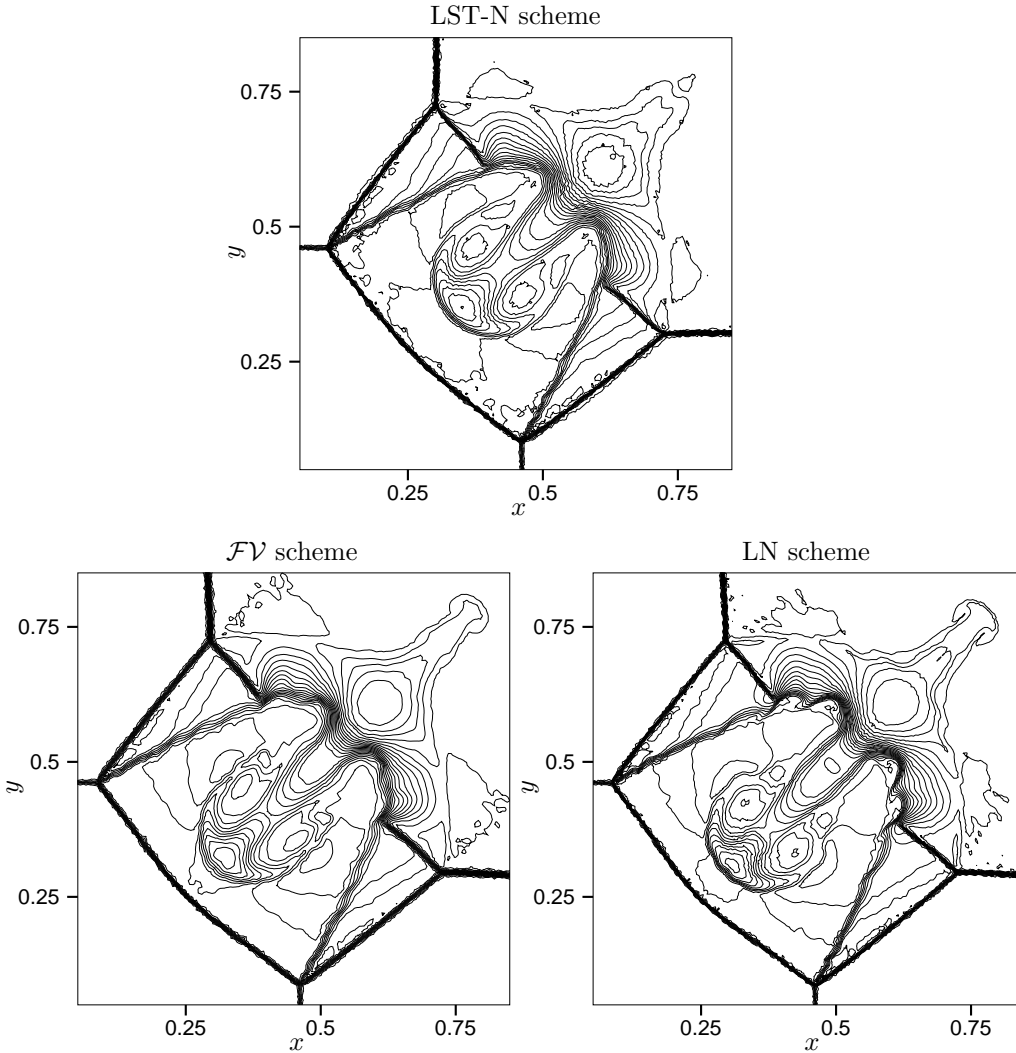


Figure 10.13: Shock-shock interaction. Contours of the density obtained with the LST-N scheme (top), \mathcal{FV} scheme (bottom-left) and LN scheme (bottom-right)

The plots in figure 10.13 show that, roughly, all the schemes capture the complex structures in the flow. However, the resolution of the interaction is much better reproduced in the pictures on the bottom, the LST-N giving by far the worst results. This can be seen from the poor resolution of integration between the slip lines emanating from the lower triple point and of the jet along the diagonal. Conversely, these features are very well captured by the \mathcal{FV} scheme and by the LN scheme. The latter, in particular, gives a very rich solution, in which the formation of KH instabilities, in correspondence of the contact lines interacting with the upper lambda-shock is already visible. As for the double Mach reflection, the LN scheme gives the best result.

To confirm this, we plot on figure 10.14 the distributions of the density and of the pressure along the diagonal, computed by the \mathcal{FV} scheme and by the LN scheme. The latter clearly resolves better the compression of the fluid across the jet. We also remark that both the pressure and density profile are absolutely free of numerical oscillations. The profiles obtained with the LST-N scheme, which are not reported on the plots for clarity, confirm the poor resolution of the jet which, as one can see on figure 10.13, is much weaker and slower.

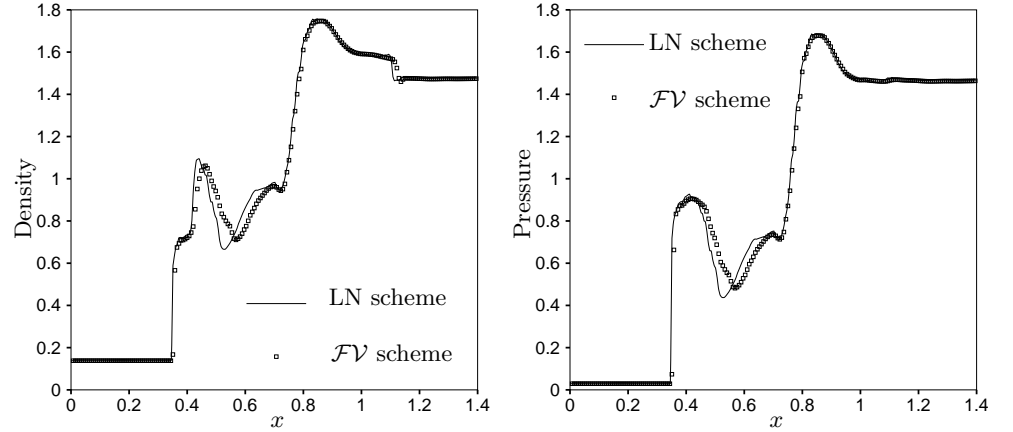


Figure 10.14: Shock-shock interaction. Density (left) and pressure (right) distribution along the symmetry line. \mathcal{FV} scheme (symbols) and LN scheme (line)

The reasons of this poor accuracy are not completely understood. As the results of §10.2.2 show, it seems that the limiting technique used here leads to schemes qualitatively insensitive to the nature of the underlying linear scheme across single discontinuities. However, in presence of more complex structures, it is apparent that this is not true anymore. This is also confirmed by the scalar results of §7.3.1 and §7.3.2. The limiting technique guarantees excellent shock capturing. However, the accuracy of the limited scheme is a different matter. The \mathcal{LP} property not being questioned, poor accuracy could mean that some kind of weak instability is present. Hence, once more, we are led to the conclusion that the understanding of the stability properties of these nonlinear schemes is the most important missing piece of the construction.

10.2.4.1 Grid refinement with the LN scheme

We present here the results of a grid refinement study performed with the LN scheme. We computed this test-case on three meshes with sizes $h = 1/100$, $h = 1/200$ and $h = 1/400$. The results are visualized on figure 10.15 in terms of density contours.

Despite of the irregularity of the grid, the solutions are quite clean and the approximation of the shocks converges in a very stable manner. The formation and break-up of the KH instability in the fine mesh solution is clearly visible. One can refer to [120] for similar results obtained with a nonlinear space-time scheme based on the use of the conservative linearization, and to [60, 61] for results obtained with a different nonlinear \mathcal{RD} scheme on structured triangular meshes.

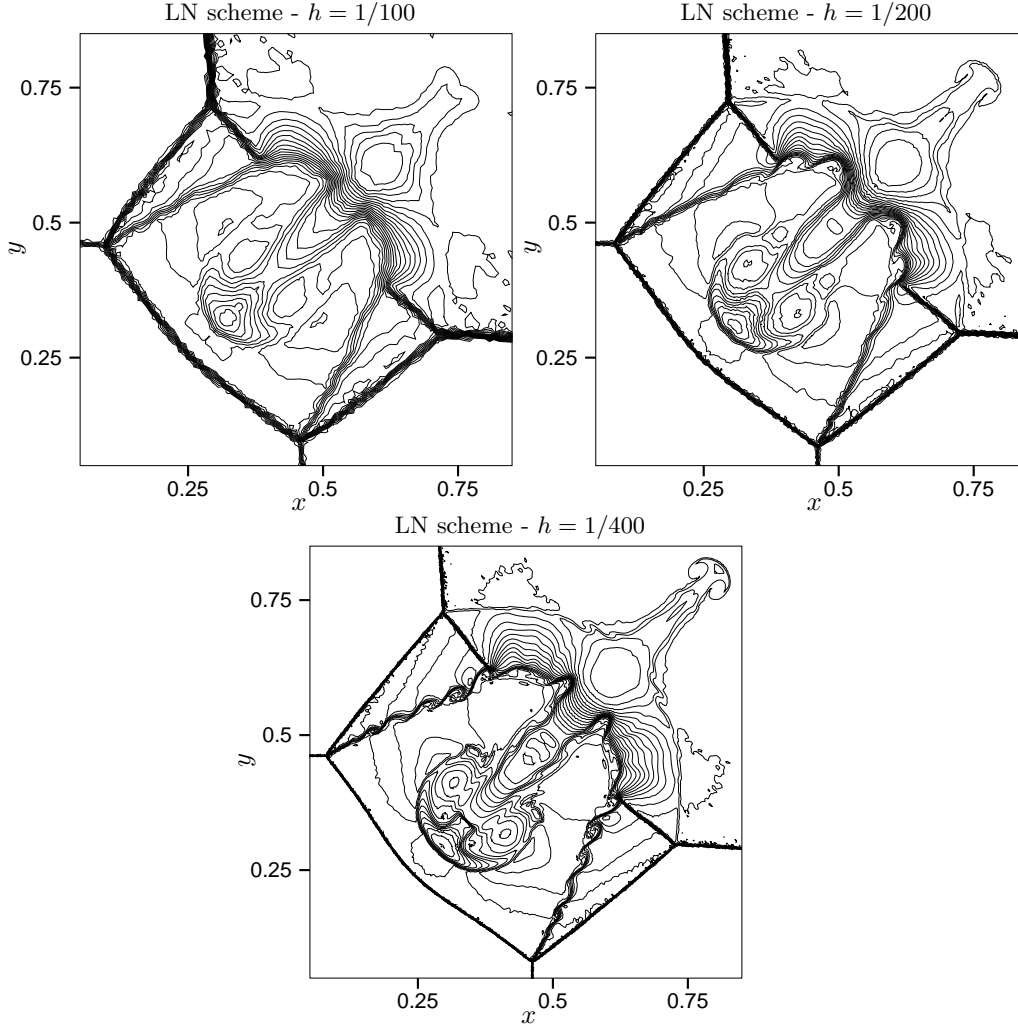


Figure 10.15: Shock-shock interaction: grid refinement with the LN scheme. Contours of the density. Top-left: $h = 1/100$. Top-right: $h = 1/200$. Bottom: $h = 1/400$

10.2.5 A shock-bubble interaction

At last, we test our schemes on a problem involving the interaction of a shock with a steady contact discontinuity: a shock-bubble interaction. The problem consists of the interaction of a Mach 2.95 right-moving shock with a *spot* of hot gas. A sketch of the initial state is given on the top-left on figure 10.16. This is a well known test-case [88, 107]. A thorough description of the problem with extensive numerical results obtained using the front tracking method of [88] is also available on-line¹. In particular, the spatial domain is the rectangle $[-0.1, 1.5] \times [-0.5, 0.5]$. At time $t = 0$ the shock is positioned at $x = 0$. The steady circular discontinuity in density and temperature is initially centered at $x = 0.3$ and $y = 0$. The density inside this *bubble* is $\rho = 0.1$. As in [88], only half of the domain has been modeled, setting a symmetry boundary condition along the line $y = 0$. The problem has been solved with the LST-N and LN schemes on an irregular mesh with $h = 1/200$. The results at times $t = 0.1$, $t = 0.15$ and $t = 0.25$ are presented in terms of density contours in figure 10.16. Note that in all the pictures we have plotted on the bottom the solutions obtained with the LN scheme, and on the top the ones computed by the LST-N scheme.

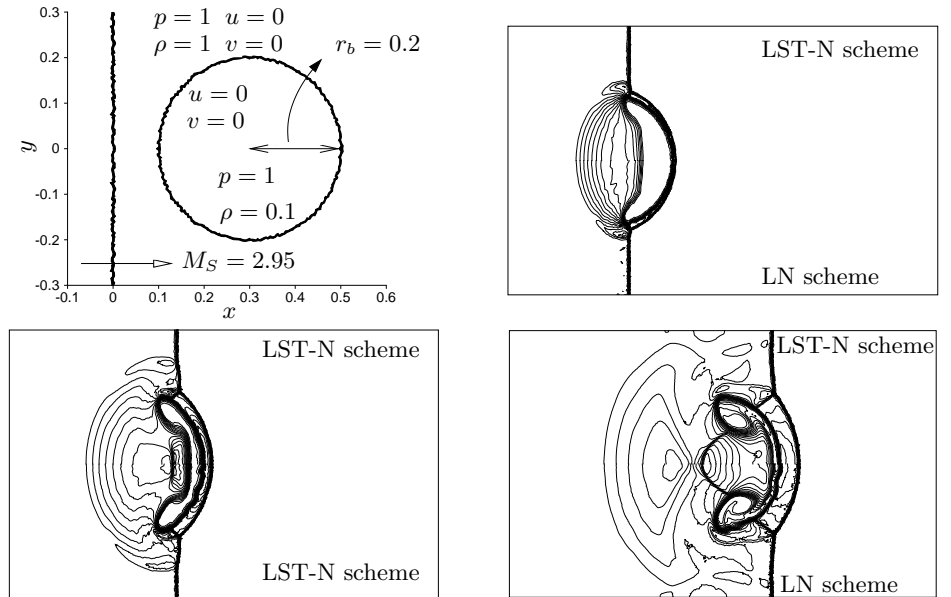


Figure 10.16: Shock-bubble interaction. Initial solution (top-left) and density contours at $t = 0.1$ (top-right), $t = 0.15$ (bottom-left) and $t = 0.25$ (bottom-right). Top half of the plots: LST-N scheme. Bottom-half of the plots: LN scheme

¹<http://www.math.ntnu.no/~andreas/fronttrack/gas/sb/>

As a result of the interaction, the incoming shock is partially transmitted into the hot fluid and partially reflected as an expansion, while the contact itself is set into motion. At time $t = 0.15$ (bottom-left in figure 10.16) the transmitted shock has already crossed the right boundary of the circular discontinuity which now starts folding into a well known symmetric structure containing two rollers. This is clearly seen at time $t = 0.25$ (bottom-right in figure 10.16). Compared to the results of [88, 107], computed on structured meshes, the interaction is well reproduced by both schemes. However, in the solution obtained with the LN scheme the roll-ups are crisper and definitely better resolved. Also, in the solutions of the LN scheme at times $t = 0.15$ and $t = 0.25$ the interface of the contact has a *wavy* pattern giving the *glimpse* of an inviscid instability.

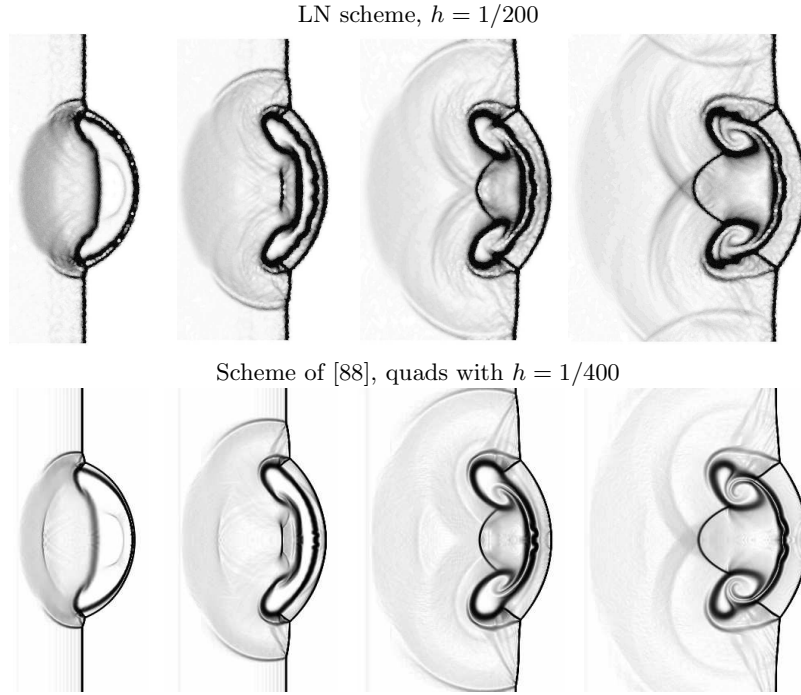


Figure 10.17: Shock-bubble interaction. Numerical Schlieren visualizations. From left to right: $t = 0.10$, $t = 0.15$, $t = 0.20$ and $t = 0.25$. Top row: LN scheme, irregular grid with $h = 1/200$. Bottom row: Scheme of [88], structured grid with $h = 1/400$.

To further confirm the quality of our results, we compare on figure 10.17 numerical Schlieren visualizations of the solution of the LN scheme with the ones of [88], obtained on a structured grid of quads with $h = 1/400$, and available on-line¹. The Schlieren visualizations of our results have been obtained following the procedure described in [135]. On the irregular grid with reference element size $h = 1/200$ used here, the LN scheme gives a very rich resolution of the interaction. The result is certainly comparable with the reference one, obtained on a much finer structured grid.

¹<http://www.math.ntnu.no/~andreas/fronttrack/gas/sb/>

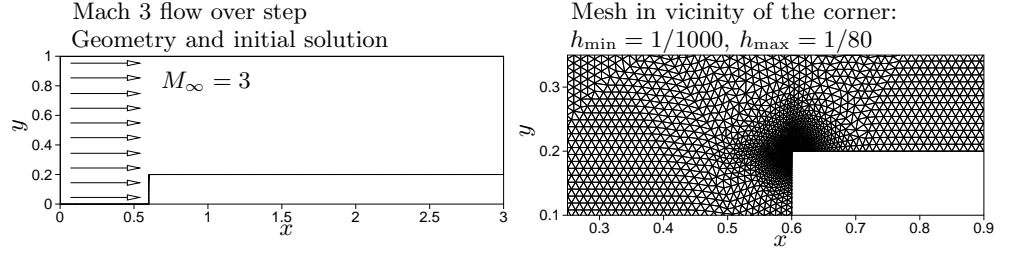
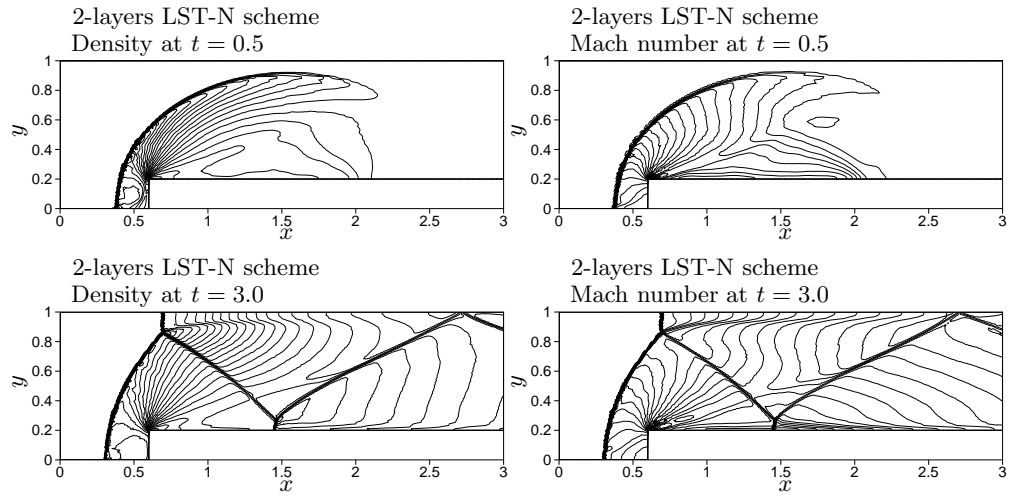


Figure 10.18: Mach 3 flow over a step. Initial solution (left) and mesh (right)

10.2.6 Two-layer schemes: Mach 3 flow over a step

For completeness, we include some results obtained with the 2-layers formulation of the *CRD* LST-N scheme (see §9.2. §7.1.5 and [8, 51]). In this section we consider the well known problem of a Mach 3 flow over a forward facing step of [187]. The geometry of the spatial domain is sketched on the left on figure 10.18. The initial solution consists of a uniform Mach 3 flow. The problem has been solved on an unstructured triangulation topologically different from the one reported on figure 3.1. For this reason, on the right on figure 10.18, we show a close-up of the mesh in vicinity of the corner. Far from this point, this grid is somewhat more regular than the ones used for the other tests, and has a reference element size $h_{\max} = 1/80$. At the corner, h reduces to $h_{\min} = 1/1000$.


 Figure 10.19: Mach 3 flow over a step. Density (left) and Mach number (right) contours at time $t = 0.5$ (top row) and $t = 3.0$ (bottom row). Results computed using the 2-layers LST-N scheme with $Q = \Delta t_2 / \Delta t_1 = 10$

As remarked in §9.2.1, when using the 2-layers formulation of the schemes, the time-step Δt_2 in the second layer can be chosen in different ways. As in [8] and [53], due to the presence of the refined region close to the corner, we have fixed the ratio between the time-steps in the first and second layer to $Q = 10$ (see equation (9.3) in §9.2.1). The results obtained with the 2-layers LST-N scheme are reported on figure 10.19 in terms of density and Mach number contours at times $t = 0.5$ and $t = 3.0$. The contour plots show the narrow and non-oscillatory approximation of the curved shock and of its reflections. The slip line emanating from the triple point is also well resolved. Similar results have been obtained in [8] with the 2-layers LN scheme based on the conservative linearization, and in [47, 53] with a 2-layers space-time blended scheme.

10.2.7 Two-layer schemes: slow shock hitting a wedge

We consider the interaction of a slow shock with a wedge [8, 191]. The spatial domain is the rectangle $[-1, 2.5] \times [-1.1, 1.1]$. At time $t = 0$ a right moving Mach 1.5 shock is placed at $x = -0.5$. At $x = 0$ is positioned a wedge of length and height 0.5 (see sketch on figure 10.20). For symmetry reasons, only the top half of the interaction has been simulated, setting symmetry BCs along the line $y = 0$. The computations have been run with the 2-layers LST-N scheme on an irregular triangulation with $h = 1/100$. For this test, we have fixed the magnitude of the total time-step $\Delta t = \Delta t_1 + \Delta t_2 = 0.005$, leading to values of $Q = \Delta t_2/\Delta t_1$ between 5.406 and 5.65.

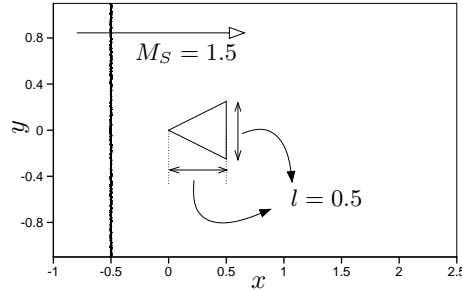


Figure 10.20: Slow shock hitting a wedge. Sketch of the initial solution

Contour plots of the computed density at $t = 0.6$, $t = 0.7$, $t = 0.95$ and $t = 1.65$ are shown on figure 10.21. Experimental Schlieren visualizations of the interaction, taken from [180], are also reported for a *qualitative* comparison¹. The reflection of the shock and its diffraction around the trailing corner are well resolved. The spurious entropy generated in correspondence of the corner causes the formation of vorticity, which can be seen on the bottom pictures. This generation of vorticity, in reality related to viscous effects, is typical of shock capturing schemes. On the bottom picture, we can see the interaction of the diffracted shock with these *hot spots* of vorticity. The results are comparable to the ones of [191].

¹no quantitative details are given in [180] on the geometry of the wedge and on the strength of the slow incoming shock

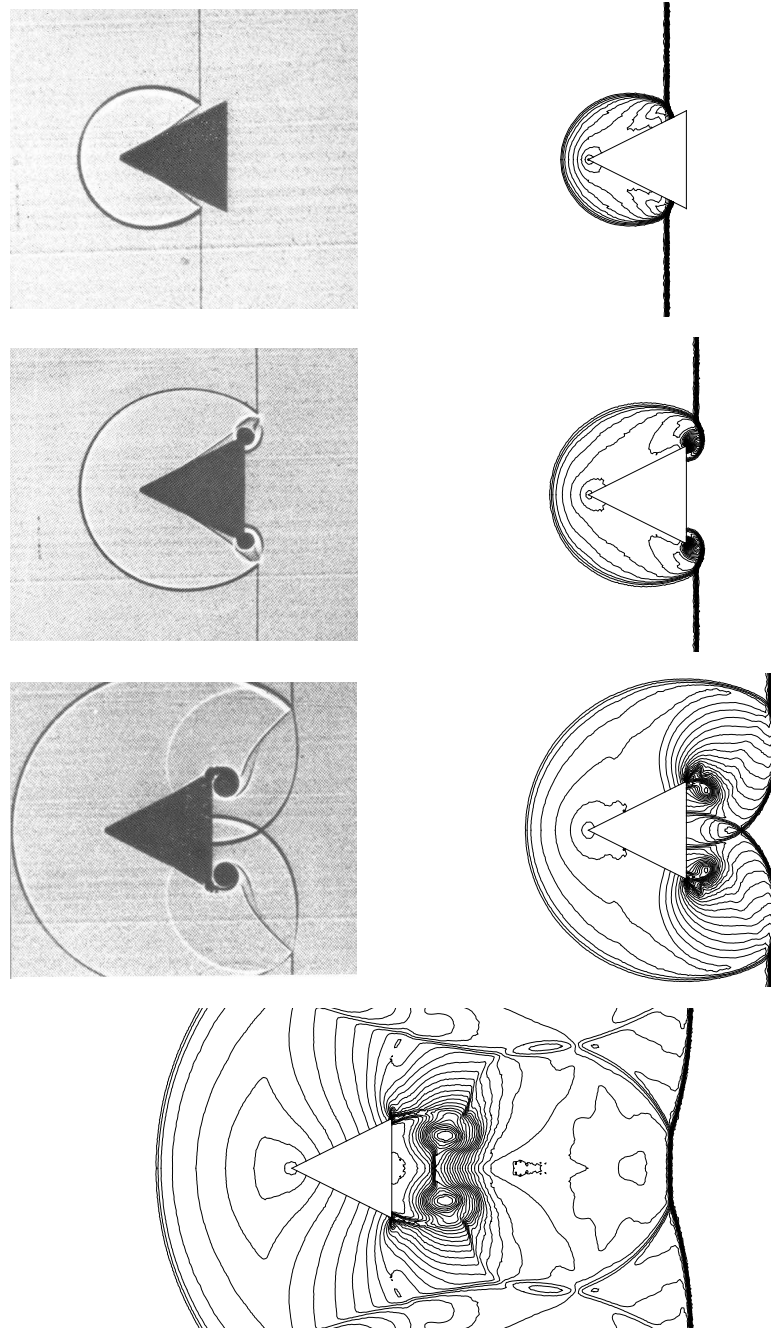


Figure 10.21: Slow shock hitting a wedge. Experimental Schlieren visualizations [180] (left column) and contours of the computed density field (right column and bottom). Numerical solutions: from top to bottom $t = 0.6$, $t = 0.7$, $t = 0.95$ and $t = 1.65$. Results obtained using the 2-layers LST-N scheme with $\Delta t = \Delta t_1 + \Delta t_2 = 0.005$

10.3 Summary

This chapter has presented a large database of results obtained by solving the Euler equations of a perfect gas with the \mathcal{CRD} schemes. The main conclusions we can draw from these results are the following.

- The \mathcal{CRD} N scheme and its space-time variants are indeed very stable and robust. The solution of very difficult problems on irregular meshes has shown their ability to resolve discontinuities in a non-oscillatory way;
- The limiting procedure allows to construct nonlinear schemes with excellent shock capturing and monotonicity preserving properties;
- Shock capturing and linearity preservation being guaranteed by construction, the limiting technique seems to have solved the problem of the design of nonlinear high-order schemes. However, the results obtained with the LST-N scheme on some of the tests confirm the lack of accuracy already observed in §7.3.1. This might be an hint to the presence of weak instabilities. The understanding of this aspect is very limited and needs to be improved;
- The space-time LN scheme has proved very competitive, in terms of accuracy on irregular grids, with \mathcal{FV} , WENO and \mathcal{DG} schemes. The main weakness of the \mathcal{RD} approach remains perhaps its cost, due to the implicit character of the schemes and to the considerable number matrix operations needed to compute the local nodal residuals;
- Results with the 2-layers variant of the \mathcal{CRD} LST-N scheme have shown that the quality of the results obtained with the \mathcal{RD} approach is retained by this double layer formulation, while giving the advantage of an unconditionally monotone time marching procedure.

Chapter 11

A two-phase flow model

In this chapter we consider the solution of the system of \mathcal{CL} s defined by the following set of conserved variables and fluxes

$$\mathbf{u} = \begin{bmatrix} \alpha_g \rho_g \\ \alpha_l \rho_l \\ \rho u \\ \rho v \end{bmatrix}, \quad \mathcal{F}(\mathbf{u}) = \begin{bmatrix} \alpha_g \rho_g u & \alpha_g \rho_g v \\ \alpha_l \rho_l u & \alpha_l \rho_l v \\ \rho u^2 + p & \rho uv \\ \rho uv & \rho v^2 + p \end{bmatrix} \quad (11.1)$$

where α_g and α_l are the gas and liquid *volume fractions*, ρ_g and ρ_l are gas and liquid densities, $\vec{u} = (u, v)$ is the local flow speed, ρ is the mixture density

$$\rho = \alpha_g \rho_g + \alpha_l \rho_l. \quad (11.2)$$

and p is the pressure. The model is closed by the relation

$$\alpha_g + \alpha_l = 1. \quad (11.3)$$

and by the EOS relating the densities to the pressure. In the following we will denote by α the gas volume fraction, assuming implicitly that α_l is obtained from (11.3). We will often refer to α as to the *void* fraction. Concerning the EOS, we have used as in [128] the following relations representative of air and water (S.I. units are used):

$$p = \Gamma_g \left(\frac{\rho_g}{\rho_{g0}} \right)^{\gamma_g} \quad (11.4)$$

with $\Gamma_g = 10^5$, $\rho_{g0} = 1$, $\gamma_g = 1.4$, and

$$p = \Gamma_l \left[\left(\frac{\rho_l}{\rho_{l0}} \right)^{\gamma_l} - 1 \right] + p_{l0} \quad (11.5)$$

with $\Gamma_l = 3.31 \times 10^8$, $\rho_{l0} = 1000$, $\gamma_l = 7.15$, 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, the model is in \mathcal{CL} form and 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 11.1, on the right picture in the same figure we plot the pressure, void fraction and x -velocity ratios as functions of the Mach number

$$M_R = \frac{u_R}{\sqrt{p_R/\rho_R}} \quad (11.6)$$

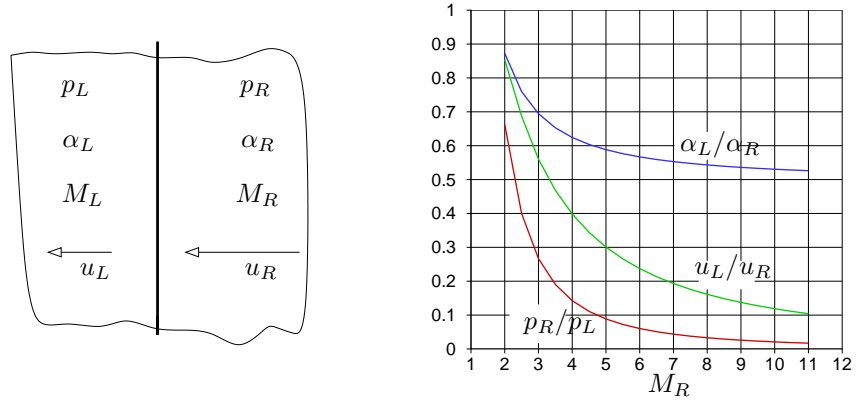


Figure 11.1: Jump conditions for the two-phase model. Flow from right to left.

Across the shock, while the pressure increases, the gas volume fraction decreases. This is a consequence of the higher compressibility of the gas, whose density increases more rapidly with the pressure, hence leading to a smaller specific volume with respect to the liquid phase. 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. 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 (11.3), a nonlinear equation for the pressure is obtained which can be solved in a few Newton iterations (see [128] for more). In conclusion, even being so simple, this model has all the features of systems of conservation laws with complex thermodynamics. The application of our schemes to these equations will prove their flexibility and further confirm their robustness.

11.1 Time-dependent computations

11.1.1 Moving shocks in air-water mixtures

We start by considering the computation of a planar shock moving in a quiescent two-phase mixture containing 50% gas and 50% liquid ($\alpha_{lR} = \alpha_{gR} = 0.5$) at a pressure $p_R = 10^6$. The shock Mach number is set to $M_S = 3$. The spatial domain is the rectangle $[0, 2] \times [0, 0.1]$. As in §10.2.1, we have run the simulations on both regular and irregular triangulations with element size $h = 1/100$. Periodic boundary conditions are imposed on the top and bottom boundaries. We will discuss the results obtained on the unstructured grid, which are indistinguishable from the ones obtained on the structured one. At time $t = 0$ the shock is located at $x = 0.5$. The final time of the simulation has been set to $t_f = 1/u_S$, corresponding to a displacement of the *analytical* shock of a unit length. We present the solutions of the space-time \mathcal{CRD} N, ST-N, LN and LST-N schemes (see §9.2). The output is visualized by extracting the data along the line $y = 0.05$.

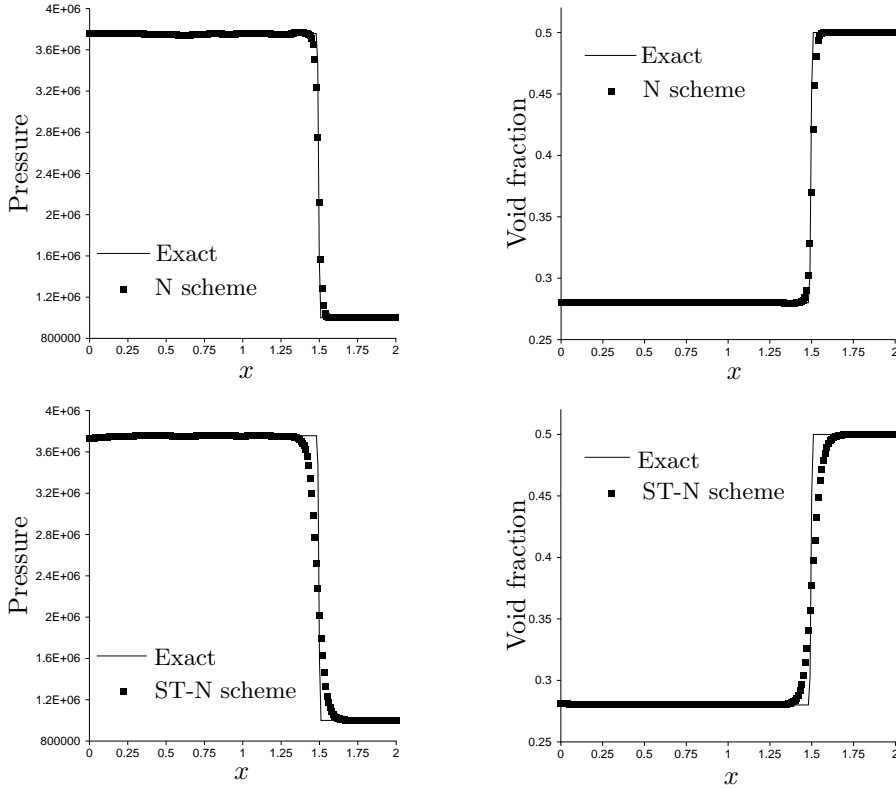


Figure 11.2: Two-phase $M_S = 3$ shock. Pressure (left) and void fraction (right) along the line $y = 0.05$. Solutions of the N (top) and ST-N (bottom) schemes

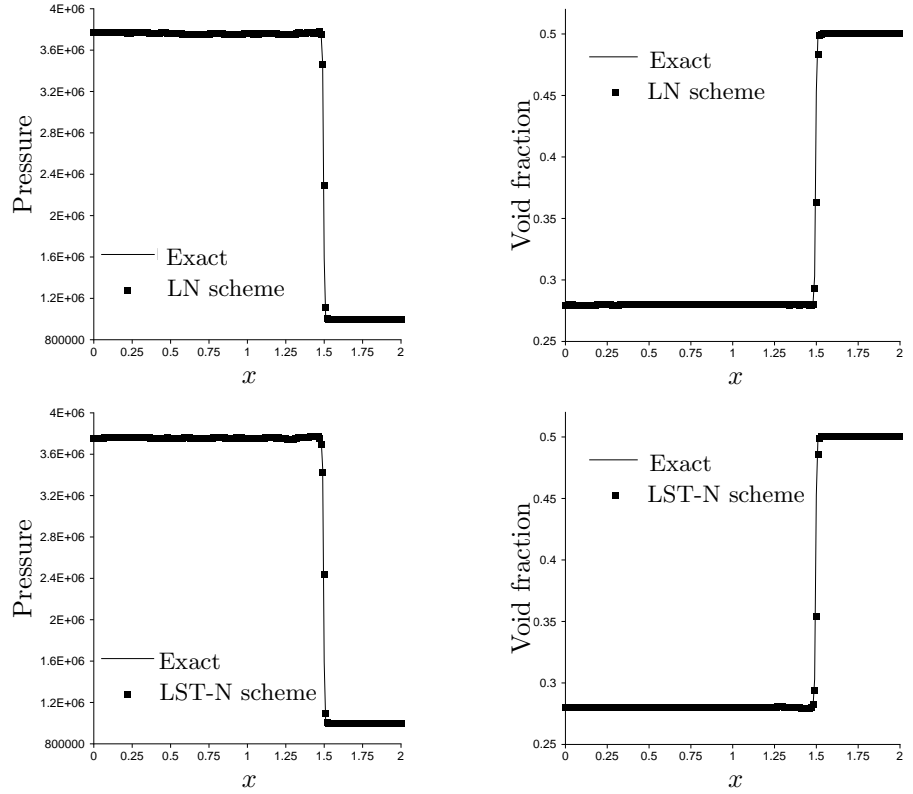


Figure 11.3: Two-phase $M_S = 3$ shock. Pressure (left) and void fraction (right) along the line $y = 0.05$. Solutions of the LN (top) and LST-N (bottom) schemes

The results are reported on figure 11.2, for the linear schemes, and figure 11.3, for the nonlinear ones. The shock position is correctly computed and the non-oscillatory character of the results is evident. The perturbations seen in the case of the Euler equations (see §10.2.1) are not visible, probably due to the weaker character of this shock. The nonlinear schemes give a very sharp and monotone capturing of the discontinuity.

11.1.2 A two-phase 2D Riemann Problem

This problem is meant to be an analog of the two dimensional Riemann problem of §10.2.2. The initial solution is given by a quiescent mixture with $\alpha = 0.5$, in which the following discontinuity in the pressure is imposed:

$$p = \begin{cases} 10^7 & \text{if } xy \geq 0 \\ 10^8 & \text{otherwise} \end{cases}.$$

The problem is solved on the domain $[-5, 5]^2$ up to time $t_f = 0.004$ on an unstructured mesh with $h = 1/10$. Symmetry BCs are imposed on all the boundaries. On figure 11.4 we plot the contours of the mixture density (11.2) corresponding to the solutions obtained with the N, ST-N, LN and LST-N schemes.

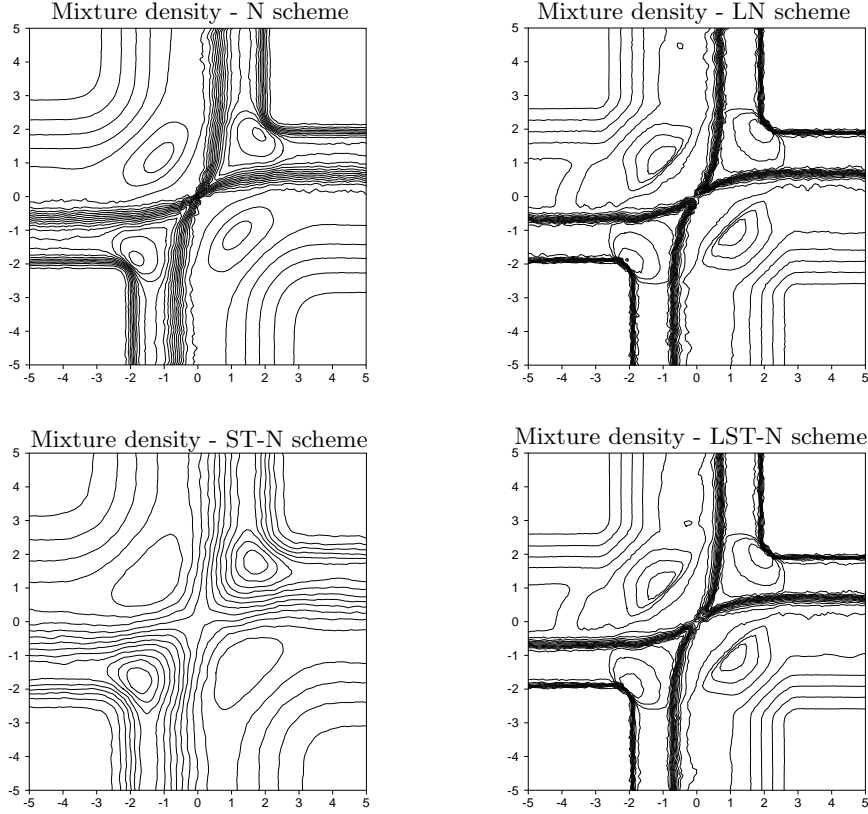


Figure 11.4: Two-phase 2D Riemann Problem. Mixture density contours at $t = 0.004$. Top: N (left) and LN scheme (right) - Bottom: ST-N (left) and LST-N scheme (right)

On the boundaries of the domain three distinct waves are visible: an expansion, a contact and a shock (left to right on the top boundary). Both the shock and the contact are noticeably better computed by the nonlinear schemes. Moving away from the boundaries, we see how the waves interact with each other. The higher resolution of the limited schemes is visible also from the fact that the lines of constant density in the expansions are kept straight for a longer distance from the boundary. As in the case of the Euler equations, there is a remarkable difference between the results of the linear ST-N and N schemes. The latter gives a visibly better resolution of the discontinuities. Nevertheless, also for this problem we see that the nonlinear LN and LST-N schemes yield nearly identical results. In figure 11.5, we compare the solutions along the top boundary of the domain with a reference, given is in this case by a 1D

numerical solution obtained on a very fine mesh containing 50000 cells with the first-order conservative \mathcal{FV} scheme of [89]. All the schemes reproduce correctly positions and strength of the shocks. The capturing of all the discontinuities is monotone and very sharp in the case of the two limited schemes.

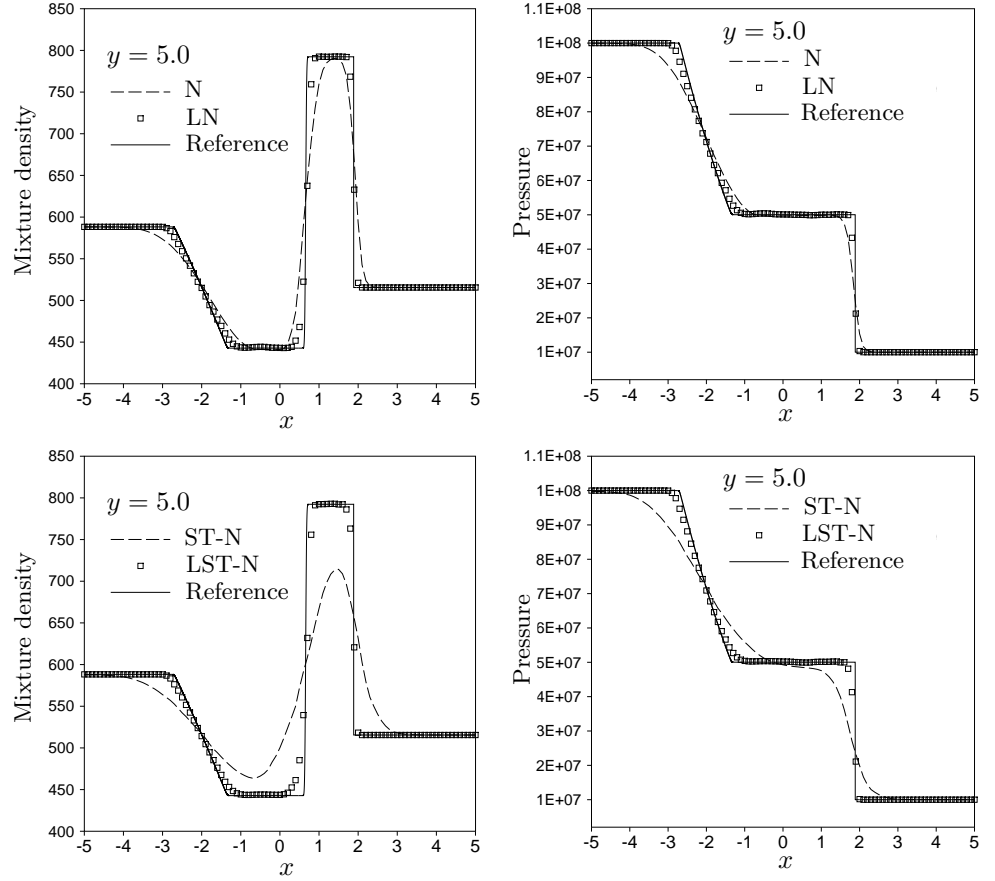


Figure 11.5: Two-phase 2D Riemann Problem. Mixture density (left) and pressure (right) at $t = 0.004$ and $y = 5.0$. Top: N and LN schemes - Bottom: ST-N and LST-N schemes

11.2 A shock-bubble interaction

The last test considered in this chapter is a two-phase *shock-bubble* interaction. The initial solution (sketched in figure 11.6) consists of a planar shock with $M_S = 3$ moving into an undisturbed quiescent mixture characterized by $\alpha_R = 0.8$ and $p_R = 10^5$. On the right of the shock we impose a stationary circular discontinuity in which the void

fraction jumps to $\alpha = 0.95$. This *bubble* is centered at $x = 0.3$ and $y = 0$, and its radius is $r_b = 0.2$. Due to the symmetry of the interaction, we have simulated only half of the problem, setting symmetry boundary conditions along the line $y = 0$. We present the results obtained with the LN and LST-N schemes on an unstructured grid with reference element size $h = 1/200$.

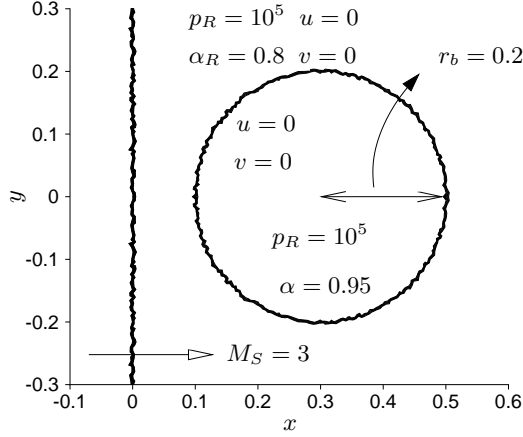


Figure 11.6: Two-phase shock-bubble interaction. Initial solution

The results are visualized in terms of contours of the mixture density (11.2) in figure 11.7. As in §10.2.5, in each picture we have plotted the solution obtained with the nonlinear LST-N scheme on the top half and the one obtained with the nonlinear LN scheme on the bottom half. From the figures we see the shock partially transmitted through the void fraction discontinuity and partially reflected as an expansion, while the contact itself is set into motion (pictures on the top row). Once the *undisturbed* shock has crossed the region occupied by the whole circular discontinuity, and joined the transmitted shock, the interface of the contact folds, rolling-up into a symmetric structure. The underlying physical mechanisms are clearly the same of the Euler computation presented in §10.2.5. Also in this computation the LN scheme shows a smaller numerical dissipation. Indeed, it gives a crisper resolution of the contact, its wavy structure (bottom pictures) again giving the glimpse of an inviscid instability. These results compare qualitatively well to the ones presented in [88, 107] for the Euler equation for a perfect gas and to the ones obtained with different two-phase flow models and numerics in [11, 69, 81]. However, as already remarked, our objective is not the simulation of two-phase flow *per-se*. The development of numerical methods for two-phase flow simulations represents in itself a whole research field. The contribution of this thesis is to provide a formulation of the \mathcal{RD} schemes that can be used in this field.

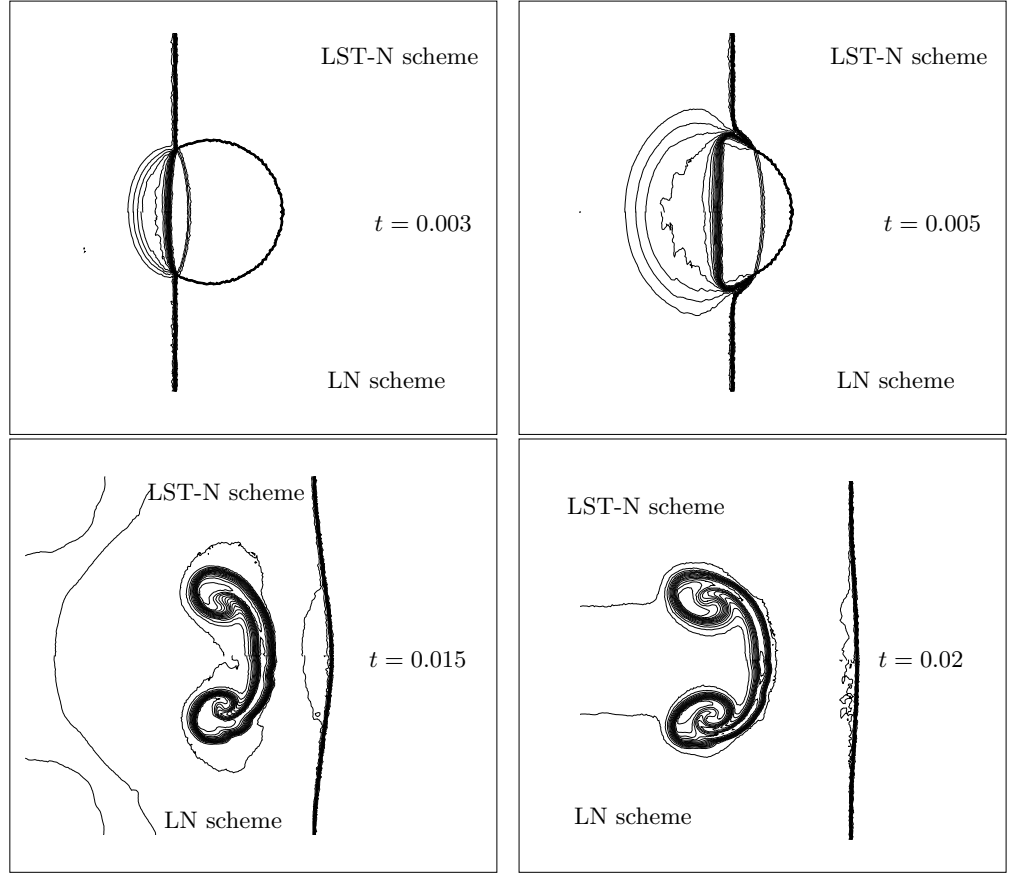


Figure 11.7: Two-phase shock-bubble interaction. Mixture density contours at times $t = 0.003$ (top-left), $t = 0.005$ (top-right), $t = 0.015$ (bottom-left) and $t = 0.02$ (bottom-right). Top half of the plots: LST-N scheme. Bottom-half: LN scheme

11.3 Summary

This chapter has shown the application of the conservative space-time schemes proposed in the thesis to the solution of a two-phase flow model. The model is very simple, however it well represents systems of conservation laws with very complex thermodynamics, due to the nonlinearity of the fluxes and of the EOS. The results of this chapter show the generality and flexibility of our approach in approximating systems of \mathcal{CL} s. The schemes have confirmed their conservative and truly non-oscillatory character. In particular, also in this general context the limiting technique provides excellent shock capturing properties. As in the previous chapter, while on simple wave solutions the LN and LST-N schemes give nearly identical solutions, in more complex situations things are different. The LN scheme confirms its higher resolution.

Note that we have not addressed important issues such as the behavior of the schemes in presence of very strong contact discontinuities, especially with with non-zero transverse velocity [11, 2]. In [50], it has been shown that, provided that the pressure is used as a primary unknown, \mathcal{CRD} schemes can resolve exactly stationary contact discontinuities, *if aligned with the mesh*. Generally, however, this is not true. A detailed investigation of this topic would be very interesting.

Chapter 12

Solution of the shallow-water equations

Frictionless shallow free surface flows under the action of the gravity force are modeled by the following system of the shallow-water equations

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{u}) = \mathcal{S}(\mathbf{u}, x, y) \quad \text{on} \quad \Omega \times [0, t_f] \subset \mathbb{R}^2 \times \mathbb{R}^+ \quad (12.1)$$

with conserved variables and fluxes given by

$$\mathbf{u} = \begin{bmatrix} H \\ Hu \\ Hv \end{bmatrix} \quad \mathcal{F}(\mathbf{u}) = \begin{bmatrix} Hu & Hv \\ Hu^2 + g\frac{H^2}{2} & Huv \\ Huv & Hv^2 + g\frac{H^2}{2} \end{bmatrix} \quad (12.2)$$

with H the *local relative water height*, $\vec{u} = (u, v)$ the flow speed and g the gravity acceleration. If not stated otherwise, we assume $g = 9.81 \text{ m/s}^2$. The source term models the effects on the flow of the shape of the bottom of the bed, and is defined as

$$\mathcal{S}(\mathbf{u}, x, y) = -gH \begin{bmatrix} 0 \\ \frac{\partial B(x, y)}{\partial x} \\ \frac{\partial B(x, y)}{\partial y} \end{bmatrix} \quad (12.3)$$

where $B(x, y)$ is the local height of the bottom (see figure 12.1). We also define

$$H_{tot} = H + B$$

the total water height. We will only consider the case in which $H > 0$. Flows with $H = 0$, which are referred to as flows over *dry bed*, will not be dealt with in this thesis.

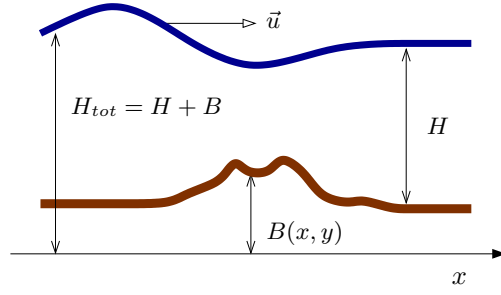


Figure 12.1: Shallow water equations: basic unknowns

As written in (12.1), these equations do not belong to the prototype of systems of \mathcal{CL} s considered up to now, due to the dependence of the source term on the relative water height H . However, using as primary variable the set of *primitive* variables

$$\mathbf{p} = \begin{bmatrix} H \\ u \\ v \end{bmatrix}$$

the system can be rewritten in the symmetric quasi-linear form

$$\frac{\partial \mathbf{p}}{\partial t} + A'_1 \frac{\partial \mathbf{p}}{\partial x} + A'_2 \frac{\partial \mathbf{p}}{\partial y} = \mathcal{S}'(x, y)$$

where the source term $\mathcal{S}'(x, y)$

$$\mathcal{S}'(x, y) = -g \begin{bmatrix} 0 \\ \frac{\partial B(x, y)}{\partial x} \\ \frac{\partial B(x, y)}{\partial y} \end{bmatrix}$$

is independent on the solution. Hence, system (12.1) can be recast into the general prototype of PDEs considered in the thesis. While being hyperbolic, the system does not admit a simple multidimensional conservative Jacobian linearization [76, 92, 132]. This makes the use of our conservative approach very well suited. However, we have to recall that, as shown in §6.2.2.2, *in the steady homogeneous case* the \mathcal{CRD} approach is a particular case of the conservative correction technique used in [92]. For completeness, we also recall that for the shallow-water equations the *Froude number*

$$Fr = \frac{\sqrt{\vec{u} \cdot \vec{u}}}{\sqrt{gH}}$$

plays the same role as the Mach number in gas dynamics.

12.1 \mathcal{RD} schemes and the lake-at-rest solution

The shallow-water equations constitute a relatively simple application for our schemes. However, they give us the chance to prove the advantage of the residual approach developed in the thesis. In particular, we recall that system (12.1) admits a class of exact steady-state solutions known as the lake-at-rest solutions. These solutions are easily obtained assuming $u = v = 0$ and integrating (12.1) over an arbitrarily small control volume \mathcal{V} obtaining

$$\int_{\mathcal{V}} \frac{\partial h}{\partial t} dx dy = - \oint_{\partial \mathcal{V}} h \vec{u} \cdot \vec{n} dl = 0 ,$$

and

$$\int_{\mathcal{V}} \frac{\partial(h\vec{u})}{\partial t} dx dy = - \int_{\mathcal{V}} gH \nabla H_{tot} dx dy .$$

If $H_{tot}(x, y, t = 0) = H_0$, $\forall(x, y) \in \Omega$, from the arbitrariness of \mathcal{V} one gets

$$\begin{aligned} H_{tot}(x, y, t) &= H_{tot}(x, y, t = 0) = H_0 & \forall(x, y) \in \Omega & \text{ and } t \geq 0 \\ u = v &= 0 & \forall(x, y) \in \Omega & \text{ and } t \geq 0 \end{aligned} \quad (12.4)$$

which defines a family of solutions parametrized by the function $B(x, y)$. Note that this is independent on the shape and regularity of $B(x, y)$, as long as ∇H_{tot} is integrable.

The analysis of §4.4.3 and §7.1.2.1 shows that as long as the numerical approximation of \mathcal{S} is second-order accurate, linearity preserving schemes will also yield second-order of accuracy. Here we are going to prove that for the shallow water equations one can do a lot better with a little extra effort.

Proposition 12.1.1 (\mathcal{LP} schemes and the lake-at-rest solution). *Denote by \mathbf{w} the set of primary variables used in the numerical approximation of (12.1). Linearity preserving \mathcal{RD} schemes preserve exactly the lake-at-rest solution, independently on topology of the mesh, regularity of $B(x, y)$ and polynomial degree of the approximation, provided that \mathbf{w} is such that for $u = v = 0$, the same continuous numerical representation is used for H and for the local height of the bottom $B(x, y)$.*

Proof. The proof is obtained quite easily by noting that with the hypotheses made, any spatial numerical reconstruction of the velocity \vec{u}_h used in the computation of the residual will be identically zero, while for H and B one has in space

$$H_h = \sum_{i \in \mathcal{T}_h} \psi_i H_i(t) \quad B_h = \sum_{i \in \mathcal{T}_h} \psi_i B_i ,$$

with i the generic node of the mesh and with the *continuous shape functions* $\psi_i(x, y)$ respecting the obvious consistency constraint

$$\sum_{j \in E} \psi_j(x, y) = 1 . \quad (12.5)$$

Consider now the the spatial residual

$$\phi^{\text{space}} = \int_E (\nabla \cdot \mathcal{F}_h - \mathcal{S}_h) \, dx \, dy .$$

The first component of ϕ^{space} is

$$\int_E \nabla \cdot (H_h \vec{u}_h) \, dx \, dy = \oint_{\partial E} H_h \vec{u}_h \cdot \hat{n} \, dl = 0 ,$$

since $\vec{u}_h = 0$. Second and third components of ϕ^{space} can be written in vector form as

$$\int_E (\nabla \cdot (H_h \vec{u}_h \otimes \vec{u}_h) + g H_h \nabla (H_h + B_h)) \, dx \, dy = \oint_{\partial E} H_h \vec{u}_h \vec{u}_h \cdot \hat{n} \, dl + g \int_E H_h \nabla H_{tot} \, dx \, dy$$

Since $\vec{u}_h = 0$, these components of the residuals reduce to

$$g \int_E H_h \nabla H_{tot} \, dx \, dy = g \int_E H_h \sum_{j \in E} H_{tot,j} \nabla \psi_j \, dx \, dy = g H_0 \int_E H_h \sum_{j \in E} \nabla \psi_j \, dx \, dy = 0$$

since on the lake-at-rest solution $H_{tot,j} = H_0 \, \forall j \in E$, and using condition (12.5). This shows that $\forall E \in \mathcal{T}_h$, $\phi^{\text{space}} = 0$ on the lake-at-rest solution, hence for any \mathcal{LP} scheme, the semi-discrete steady-state \mathcal{RD} formulation reduces to

$$\frac{d\mathbf{u}_i}{dt} = 0, \quad \forall i \in \mathcal{T}_h$$

which achieves the proof in the steady-state case. For the space-time schemes, a similar reasoning leads to the conclusion that on the lake-at-rest solution $\mathcal{F}^{n+1} = \mathcal{F}^n = 0$ (see §9.2 and equation (8.49) in §8.3.2). As a consequence, the iterative nonlinear system

$$\sum_{E \in \mathcal{D}_i} \phi_i = 0, \quad \forall i \in \mathcal{T}_h$$

reduces to the identity $0 = 0$ on this solution. A different way to see it, is to consider the explicit iterative update (9.2). One easily sees that, if in the initial solution $\vec{u} = 0$ and $H_{tot} = H_0 \, \forall (x, y) \in \Omega$, then

$$\mathbf{u}_i^{n+1,k+1} = \mathbf{u}_i^{n+1,0} = \mathbf{u}_i^0, \quad \forall i \in \mathcal{T}_h \quad \text{and} \quad \forall n \geq 0$$

□

Proposition 12.1.1 shows that differently from other numerical techniques [80, 112, 79, 93], an exact approximation of the lake-at-rest solution is achieved very naturally here, thanks to the truly residual character of the schemes. Experimentally, it is found that when setting-up an initial state with $\vec{u} = 0$ and uniform total water height, \mathcal{LP} schemes indeed preserve this state. In particular, the nodal residuals are *exactly* zero if the element residual is computed as (see equation (8.43))

$$\phi^h = \sum_{l_j=1}^3 \sum_{p=1}^{\text{NC}} \omega_p \tilde{\mathcal{F}}(\mathbf{u}_h(x_p, y_p)) \cdot \vec{n}_{l_j} + g \bar{H} \mathcal{F}_B(\mathbf{u}_h) \quad (12.6)$$

with

$$\tilde{\mathcal{F}}(\mathbf{u}) = \begin{bmatrix} Hu & Hv \\ Hu^2 & Huv \\ Huv & Hv^2 \end{bmatrix}, \quad \mathcal{F}_B(\mathbf{u}_h) = \frac{1}{2} \sum_{j \in E} \begin{bmatrix} 0 \\ (H_j + B_j)(\vec{n}_j)_x \\ (H_j + B_j)(\vec{n}_j)_y \end{bmatrix}$$

and \overline{H} the arithmetic average of the nodal values of the relative water height. Similarly, in space-time computations exact preservation of the lake-at-rest solution is achieved by computing the residual as (see equations (8.49) and (8.50))

$$\begin{aligned} \phi^h = \sum_{j \in E} \frac{|E|}{3} (\mathbf{u}_j^{n+1} - \mathbf{u}_j^n) + \frac{\Delta t}{2} \left(\tilde{\mathcal{F}}(\mathbf{u}^{n+1}) + \tilde{\mathcal{F}}(\mathbf{u}^n) \right) + \\ g \frac{\Delta t}{2} \left(\overline{H}^{n+1} \mathcal{F}_B(\mathbf{u}^{n+1}) + \overline{H}^n \mathcal{F}_B(\mathbf{u}^n) \right) \end{aligned} \quad (12.7)$$

with

$$\overline{H}^n = \frac{1}{3} \sum_{j \in E} H_j^n, \quad \overline{H}^{n+1} = \frac{1}{3} \sum_{j \in E} H_j^{n+1}$$

We will give examples of numerical computations where the use of (12.7) leads to the exact reproduction of the lake-at-rest state.

12.2 Steady-state computations

12.2.1 Super-critical flow over flat bed

This test has been performed to confirm the conservative and non-oscillatory character of the \mathcal{CRD} schemes. It consists of a super-critical $Fr = 2.74$ flow over a flat bed, in a channel with a 8.95° wedge. A sketch of the initial solution and of the geometry of the spatial domain is given on the left in figure 12.2. We have run this test with the \mathcal{CRD} N scheme and its nonlinear limited variant, the LN scheme, on an irregular mesh with $h = 1/20$. The convergence histories of the explicit iterative update (9.1) with $\nu = 0.8$ are reported on the right on figure 12.2. The general behavior is the same observed in §10.1.1: the linear scheme converges to machine accuracy without any problem, while the convergence of the limited scheme is somewhat erratic and, considering the problem (the flow is fully super-critical), relatively poor. The contours of the computed relative height and a comparison of its distribution at the outlet $x = 4$ with the exact solution are reported on figure 12.3 for both schemes. The following observations can be made. The discontinuity is captured monotonically, the nonlinear scheme giving a much sharper approximation. This is particularly clear from the plot on the bottom pictures. From the line plots we also see that angle and strength of the jump are correctly reproduced. Similar comments can be made by looking at the contour lines of the computed Froude number, (top on figure 12.4), and at its comparison with the exact solution at the outlet (bottom on figure 12.4).

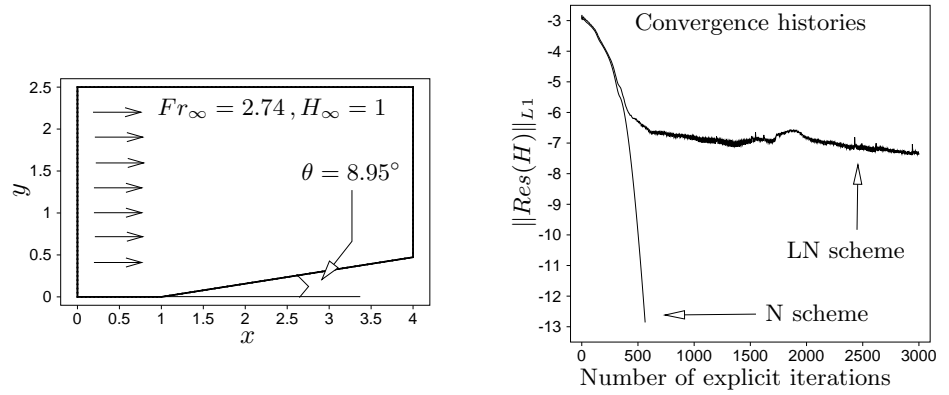


Figure 12.2: Hydraulic jump over a wedge. Sketch of the problem and convergence

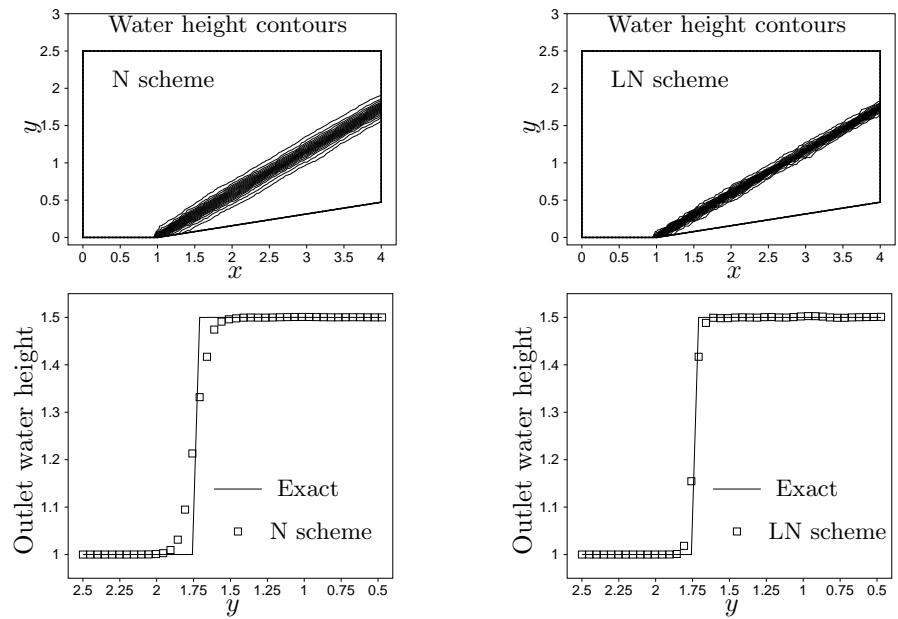


Figure 12.3: Hydraulic jump over a wedge. Water height contour levels (top) and outlet water height distribution (bottom). Results of the N (left) and LN (right) schemes

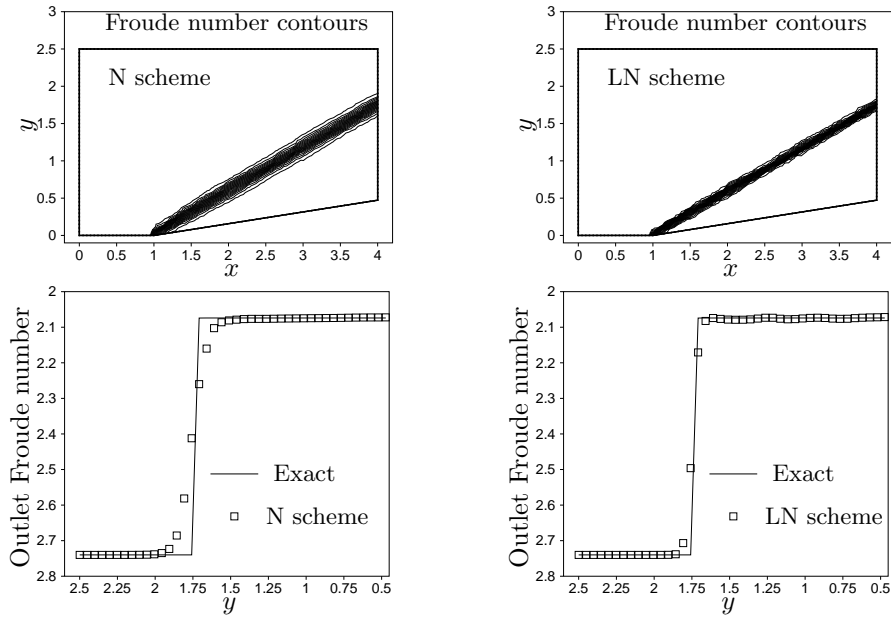


Figure 12.4: Hydraulic jump over a wedge. Froude number contours (top) and outlet Froude number distribution (bottom). Results of the N (left) and LN (right) schemes

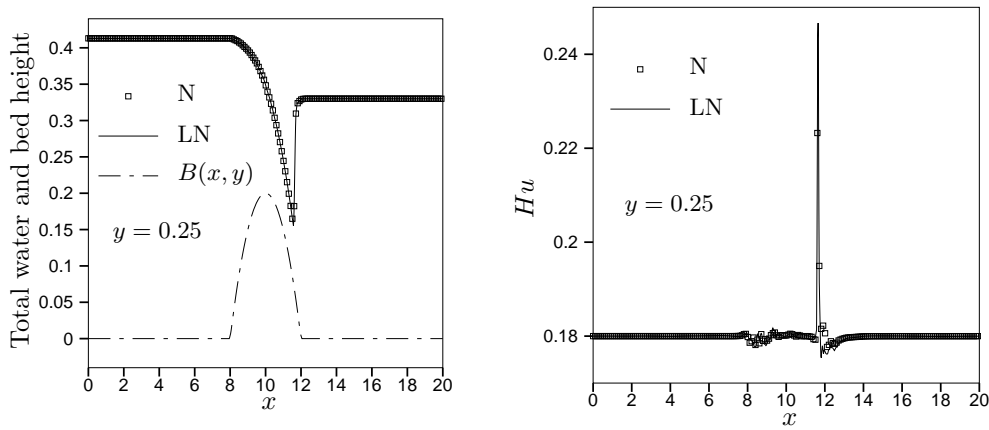


Figure 12.5: Transcritical flow over a smooth hump. Total water height (left) and discharge (right). N scheme (symbols) and limited N scheme (solid line)

12.2.2 Trans-critical flow over a smooth hump

We consider a one-dimensional test problem assuming the following variation of the bed height [153, 73, 59]

$$B(x, y) = B(x) = \begin{cases} 0.2 - 0.05(x - 10)^2 & \text{if } 8 \leq x \leq 12 \\ 0 & \text{otherwise} \end{cases} \quad (12.8)$$

Different steady solutions can be computed involving fully sub-critical, smooth trans-critical and trans-critical flow with a shock [153, 73, 59]. To assess the shock capturing capabilities of the N and LN schemes in presence of non-flat bed, we consider the case of a steady trans-critical flow with a shock. We solve the shallow-water equations on the spatial domain $[0, 20] \times [0, 0.5]$ on an irregular unstructured with reference element size is $h = 1/10$. Periodic boundary conditions are applied on the top and bottom boundaries. On the left boundary we assign the discharge $Hu = 0.18$ and zero v velocity, while on the right boundary we set $H = 0.33$. These boundaries are treated with characteristic BCs. The steady-state solutions obtained with the N scheme and the LN scheme, computing the spatial residual as in (12.6) are reported in figure 12.5, where the data along the line $y = 0.25$ are plotted. The solutions are monotone and the shock approximation is very sharp. No problems are encountered in the critical point, the acceleration being smooth in both solutions. The approximation of the discharge which should be constant and equal to 0.18 everywhere, is very good, despite of the fact that the problem has been solved on a 2D irregular mesh instead that in 1D.

12.3 Time-dependent computations

12.3.1 Break of a circular dam over flat bed

We simulate the break of a circular dam separating 10 [m] high water from a basin at $H = 0.5$ [m]. The radius of the initial discontinuity is $r = 60$ [m]. Due to the difference in water height, the flow becomes rapidly trans-critical. The computational domain is the square $[0, 100]^2$, at $t = 0$ the velocity is set to zero, while the water height is set to

$$H = \begin{cases} 10 & \text{if } r \leq 60 \\ 0.5 & \text{otherwise} \end{cases}, \quad r = \sqrt{x^2 + y^2}$$

The problem has been solved with the N, ST-N, LN and LST-N schemes on an irregular triangulation with $h = 2$. Symmetry BCs are imposed on the left and bottom boundaries. The final time of the simulations is $t_f = 3$. Contour plots of computed water height and Froude number are given in figures 12.6 and 12.7. Despite of the irregularity of the grid, the flow acceleration and the right moving water wave are well reproduced, roughly by all schemes. The contours in the region of accelerating flow are quite smooth especially with the LN scheme (top-right pictures). The capturing of the right moving water front is monotone for all the schemes. The limited schemes yield a very sharp approximation of this feature.

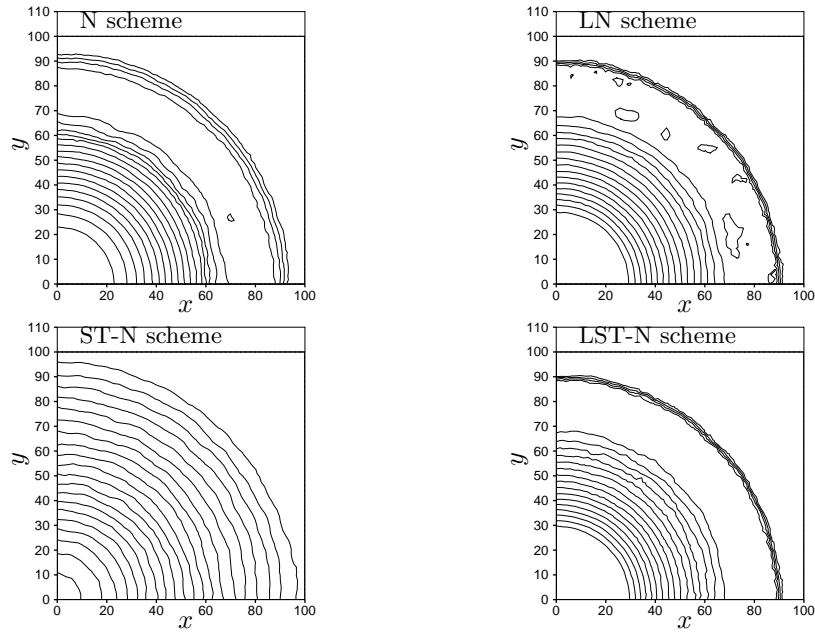


Figure 12.6: Transcritical dam-break: 20 water height contour levels at $t = 3$

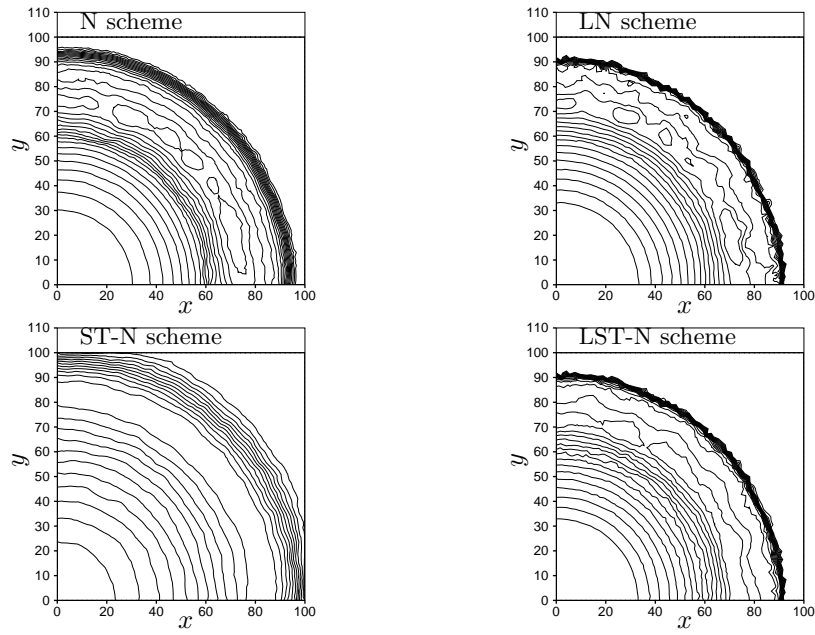


Figure 12.7: Transcritical dam-break: 20 Froude number contour levels at $t = 3$

As already noted, the ST-N scheme (bottom-left pictures) shows a considerably higher numerical dissipation compared to the N scheme (top-right pictures). Indeed, the water height waves (figure 12.6) are blurred into a unique smooth profile, while the right moving wave in the Froude contours (figure 12.7) is considerably smeared. We remark that no problem whatsoever is encountered in the critical point. To confirm the quality of the result, we show in figure 12.8 the computed water height and Froude number distributions along the line $y = x$. The figures confirm our previous observations. The flow acceleration is very smooth in all solutions and the right moving water wave is computed monotonically. The limited schemes give a sharper resolution of this last feature and the large numerical dissipation of the ST-N scheme is clear. We think these results are very good, especially the ones of the nonlinear schemes.

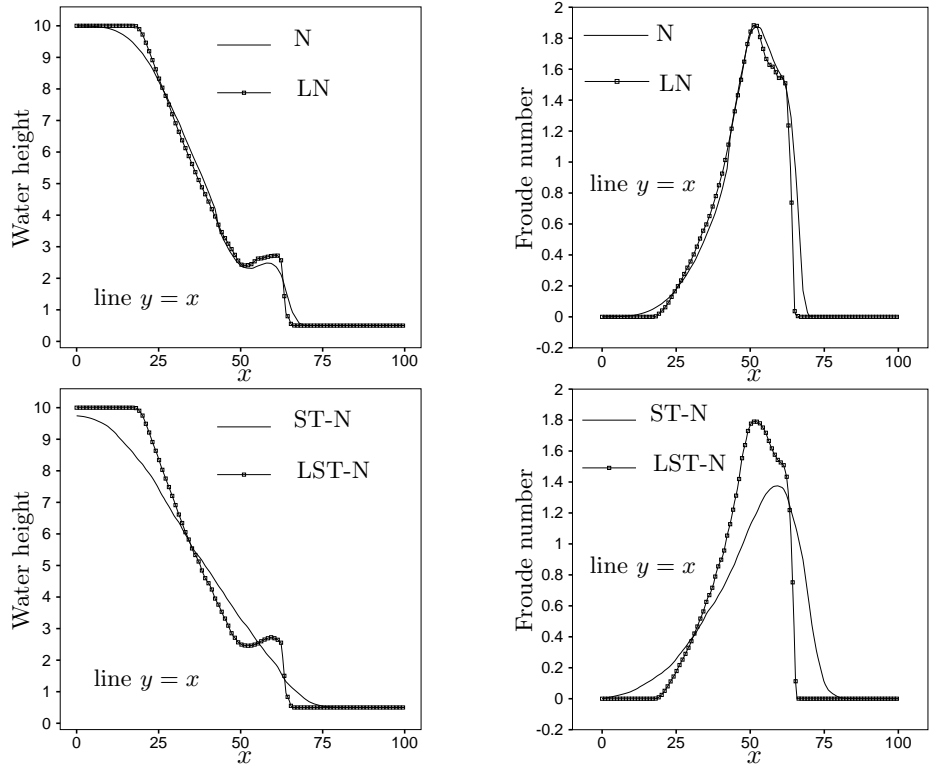


Figure 12.8: Transcritical dam-break. Water height (left) and Froude number (right) along the line $y = x$. Top: N and LN schemes. Bottom: ST-N and LST-N schemes

12.3.2 Non-symmetric dam break over flat bed

This problem is taken from [153] and is similar to the previous one, except that the geometry is more complex. We consider the sudden break of a dam separating two basins with water heights 5 and 10 [m]. The dam breaks asymmetrically at time $t = 0$ and we simulate the problem until time $t = 7.2$ [s]. The spatial domain is contained into the square $[0, 200]^2$. At $x = 95$ [m] a *breached* 10 [m] wide dam is present. The length of the breach is 75 [m] and it starts at $y = 95$. At time $t = 0$ the velocity is set to zero everywhere. A sketch of the geometry of the problem is given in figure 12.9.

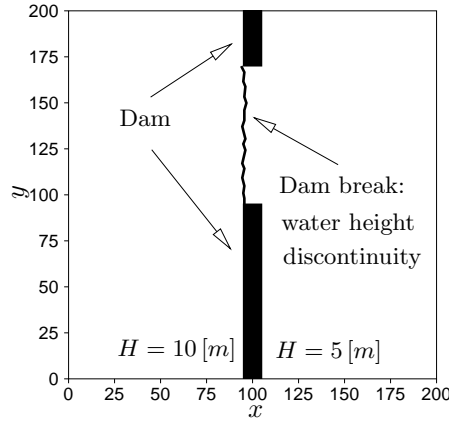


Figure 12.9: Asymmetric dam break. Problem description

The problem has been solved with the LN and LST-N schemes on an irregular grid with reference element size $h = 2$. Wall BCs are applied on all the boundaries of the domain. We show the results obtained with the limited N and limited N-ST schemes in terms of relative water height (top pictures on figure 12.10) and Froude number (bottom pictures on figure 12.10) contours. The contour plots show that both schemes compute very smoothly the water acceleration on the left of the dam, while the water wave moving to the right is very sharp and monotone in both the results. The reflection of this wave on the upper wall of the low water basin is clearly visible. We also report, on figure 12.11, the distributions of the water height and of the Froude number along the line $y = 160$. These plots confirm both the smooth character of the acceleration and the sharp and monotone capturing of the water wave. For completeness, we report on figure 12.12 a three-dimensional view of the water height at time $t = 7.2$ computed by the LN scheme.

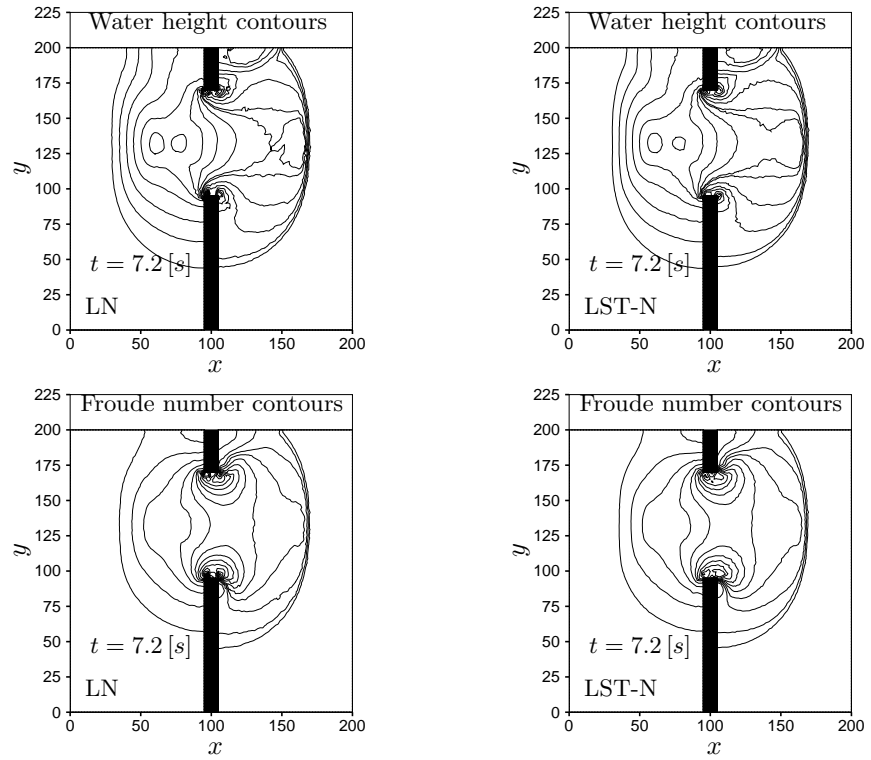


Figure 12.10: Asymmetric dam break. Water height (top) and Froude number (bottom) contours at time $t = 7.2 [s]$. LN scheme (left) and LST-N scheme (right)

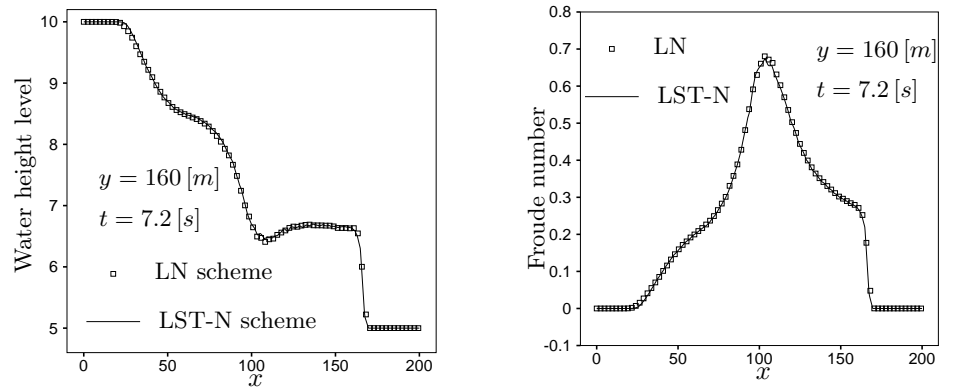


Figure 12.11: Asymmetric dam break. Water height (left) and Froude number (right) at $t = 7.2 [s]$ and $y = 160 [m]$. LN (symbols) and LST-N (solid line) schemes

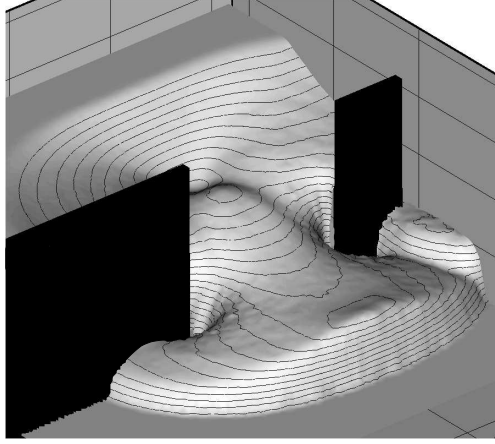


Figure 12.12: Asymmetric dam break. Water height at $t = 7.2$. LN scheme

12.3.3 Water height perturbation over smooth bed

We now consider some tests related to the approximation of the lake-at-rest solution. As stated by proposition 12.1.1, if the residual is computed according to (12.7), \mathcal{LP} schemes preserve exactly this state. This is indeed observed numerically: when initializing the solution with the lake-at-rest state, the residuals stay at machine zero throughout the computation. A class of less trivial problems that can be used to test the ability of a discretization to preserve exactly this particular solution involves initial states obtained by perturbing the exact solution. The objective of these tests is to verify that a scheme is able to resolve the evolution of the perturbation and its interaction with the non-flat bed shape, without spoiling the exact lake-at-rest state in unperturbed regions. In particular, in this section we consider a test initially proposed in [112], and more recently used in [153, 188, 189] to assess the performances of *well-balanced* formulations of very high-order relaxation, finite difference and finite volume WENO, and \mathcal{DG} discretizations. The spatial domain of the problems is $[0, 2] \times [0, 1]$. The following smooth bottom shape is assumed [112, 153, 188, 189]

$$B(x, y) = e^{-5(x-0.9)^2 - 50(y-0.5)^2}$$

corresponding to an ellipsoidal hump centered at $[0.9, 0.5]$. The initial solution is obtained by perturbing the exact lake-at-rest state in the band $x \in [0.05, 0.15]$: at $t = 0$, the velocity is set to zero everywhere, while the relative water height is set to

$$H = \begin{cases} 1.01 - B(x, y) & \text{if } 0.05 < x < 0.15 \\ 1 - B(x, y) & \text{otherwise} \end{cases}$$

We solve the problem on an unstructured discretization of the domain with reference mesh size $h = 1/100$. As in [188, 189] the gravity acceleration is set to $g = 9.812$. Characteristic BCs are imposed on the right and left end of the spatial domain, while

the upper and lower boundaries are treated as symmetry lines. We consider the solution at four different times: $t = 0.12, t = 0.24, t = 0.36$ and $t = 0.48$. In figures 12.13, 12.14, 12.15 and 12.16 we visualize the results obtained with the space-time LN scheme. In the figures, on the top rows we have reported the contours of H_{tot} computed by the LN scheme (left pictures) and the ones taken from [188] (right pictures), obtained with a fifth-order well-balanced finite difference WENO scheme. On the bottom rows, we have reported the distribution of H_{tot} along the line $y = 0.5$ and a 3D view of the solution including the bed shape. Both pictures correspond to the results obtained with the LN scheme. To obtain the 3D plots, we have reported on the third axis the scaled water height and scaled (and shifted) bed height

$$H^* = 80(H_{tot} - 1) + 1, \quad B^* = 0.5 + \frac{B(x, y)}{3} \quad (12.9)$$

The scaling of the bed height used in the line plots is reported on the figures.

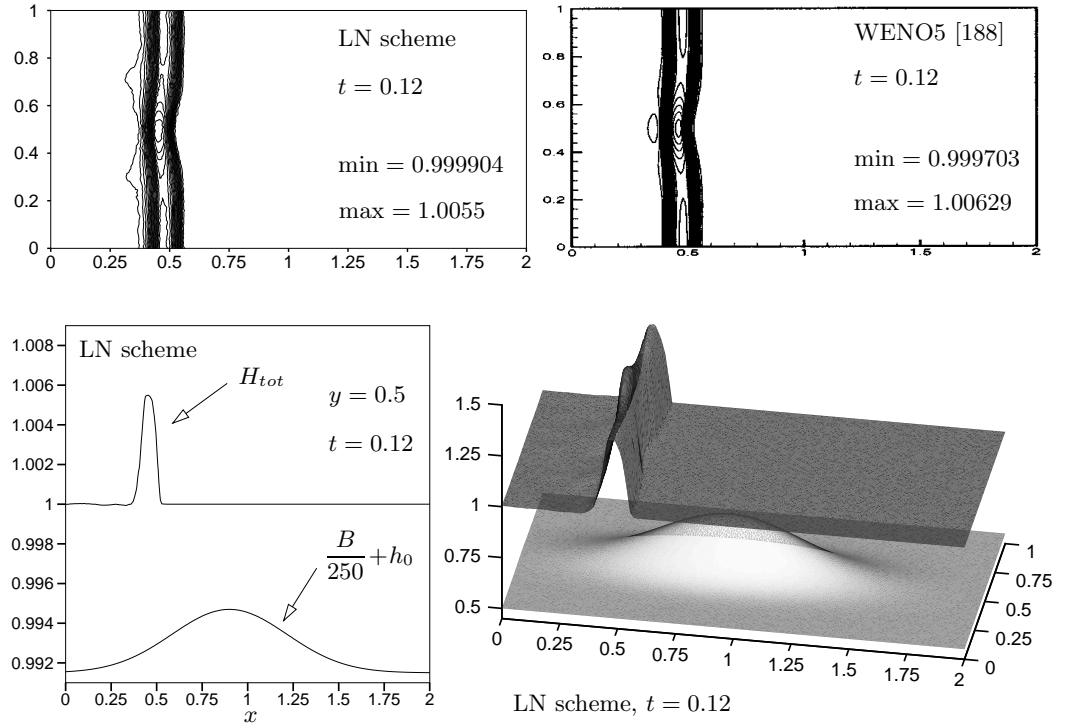


Figure 12.13: Water height perturbation over smooth bed. Solution of the LN scheme at time $t = 0.12$. Top-left: contour plot of total water height. Bottom-left: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$). Bottom-right: 3D plot of the solution. Top-right: contour plot of total water height from [188].

12.3.3. Water height perturbation over smooth bed

The following observations can be made. In the region ahead of the perturbation the exact solution is *perfectly preserved* up to machine accuracy, as predicted by proposition 12.1.1. In the region behind the perturbation the solution quickly gets back to the lake-at-rest state with a small noise, probably due to grid irregularities. With respect to the results of [188], obtained with a fifth-order finite difference WENO scheme on a structured mesh with $h = 1/100$, our results well reproduce the interaction. The small structures contained in the reference solution are visible in the results of the LN scheme. The higher accuracy of the scheme used in the reference is certainly visible from the sharper resolution of the water front, as well as from the higher (resp. lower) values of the water height in the peaks generated from the interaction. Obviously, the use of a very high-order discretization is beneficial when approximating this type of problem, involving the propagation of a small perturbation. Nevertheless, the LN scheme well reproduces the qualitative structure of the solution, while yielding a monotone approximation, and preserving exactly the lake-at-rest state in the unperturbed region.

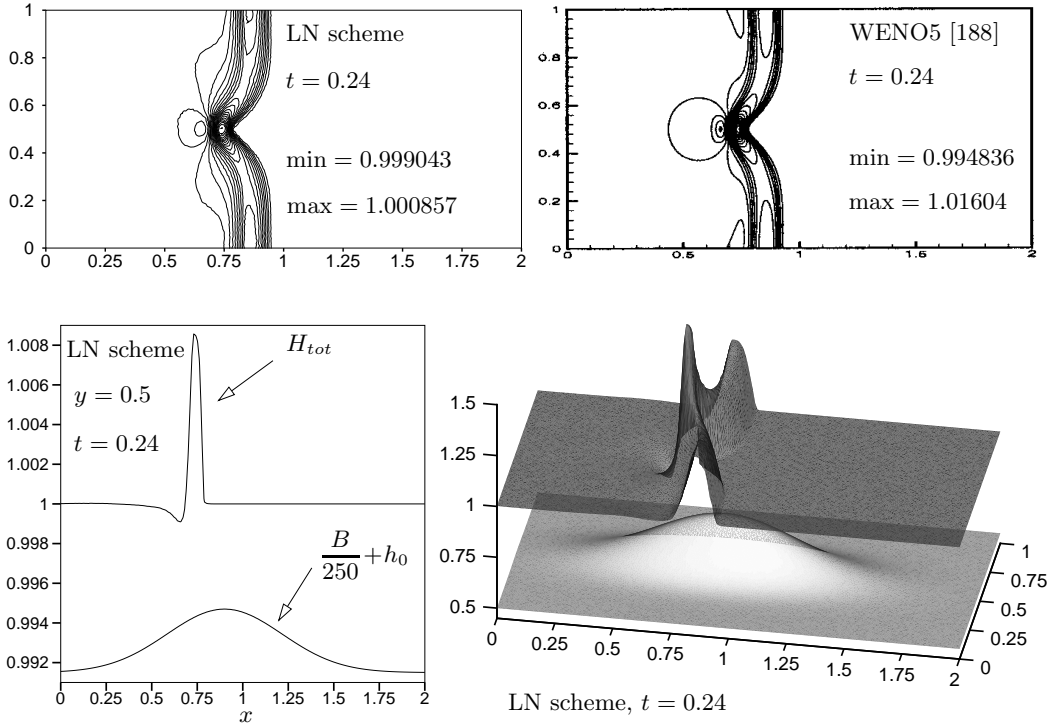


Figure 12.14: Water height perturbation over smooth bed. Solution of the LN scheme at time $t = 0.24$. Top-left: contour plot of total water height. Bottom-left: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$). Bottom-right: 3D plot of the solution. Top-right: contour plot of total water height from [188].

In particular, while this last property is a natural consequence of the residual approach used in this thesis, the well-balanced schemes of [188, 189] are based on ad-hoc constructions allowing to achieve, in the WENO and \mathcal{DG} frameworks, the exact balance between flux divergence and source term. We also add that proposition 12.1.1 applies independently on the degree of the polynomial interpolation underlying the \mathcal{RD} discretization. Very high-order \mathcal{RD} discretization guaranteeing the exact preservation of the lake-at-rest state could be constructed following *e.g.* the approach of [12, 9]. In this perspective, the results of this section are very encouraging. For completeness, we report in figure 12.17 the results obtained with the space-time N scheme at time $t = 0.6$. We clearly see the extra numerical dissipation in the smearing of the front of the perturbation. However, no oscillations are present in the solution, and in the unperturbed region the deviation from the lake-at-rest state is absolutely negligible. This is confirmed by the plot in figure 12.18, where we report the total water height and the Froude number along the line $y = 0.5$, in the unperturbed region $x \in [0.9, 2]$ at time $t = 0.6$. The plots show the preservation of the exact solution up to machine accuracy obtained with the \mathcal{LP} scheme, and the very small deviation of the N scheme.

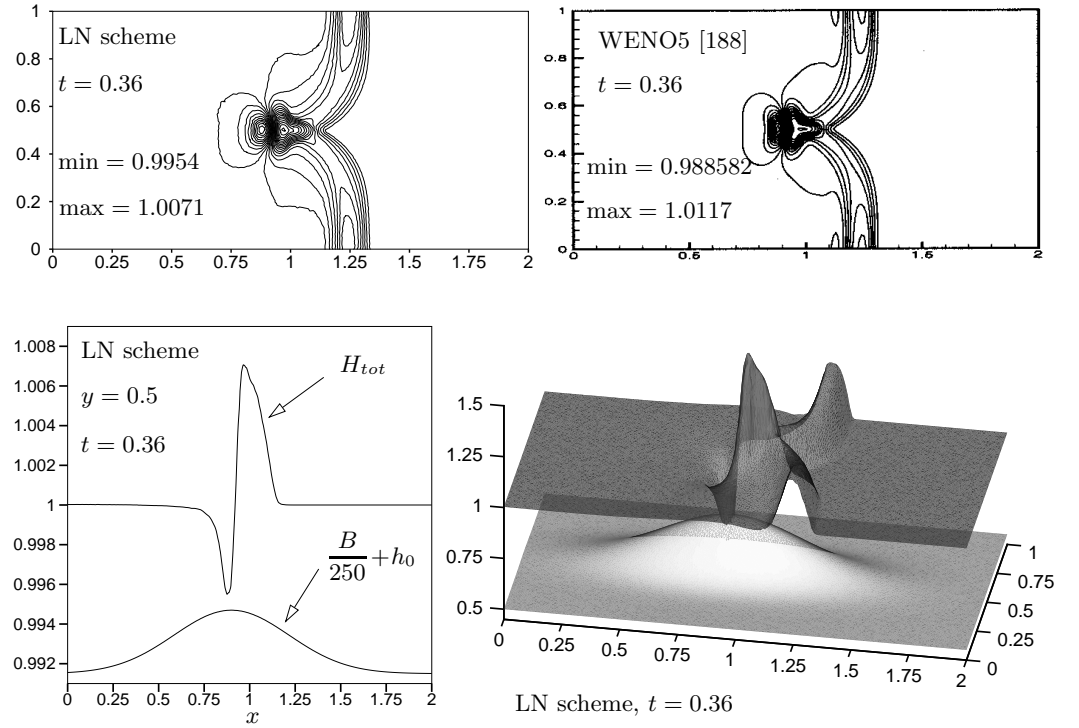


Figure 12.15: Water height perturbation over smooth bed. Solution of the LN scheme at time $t = 0.36$. Top-left: contour plot of total water height. Bottom-left: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$). Bottom-right: 3D plot of the solution. Top-right: contour plot of total water height from [188].

12.3.3. Water height perturbation over smooth bed

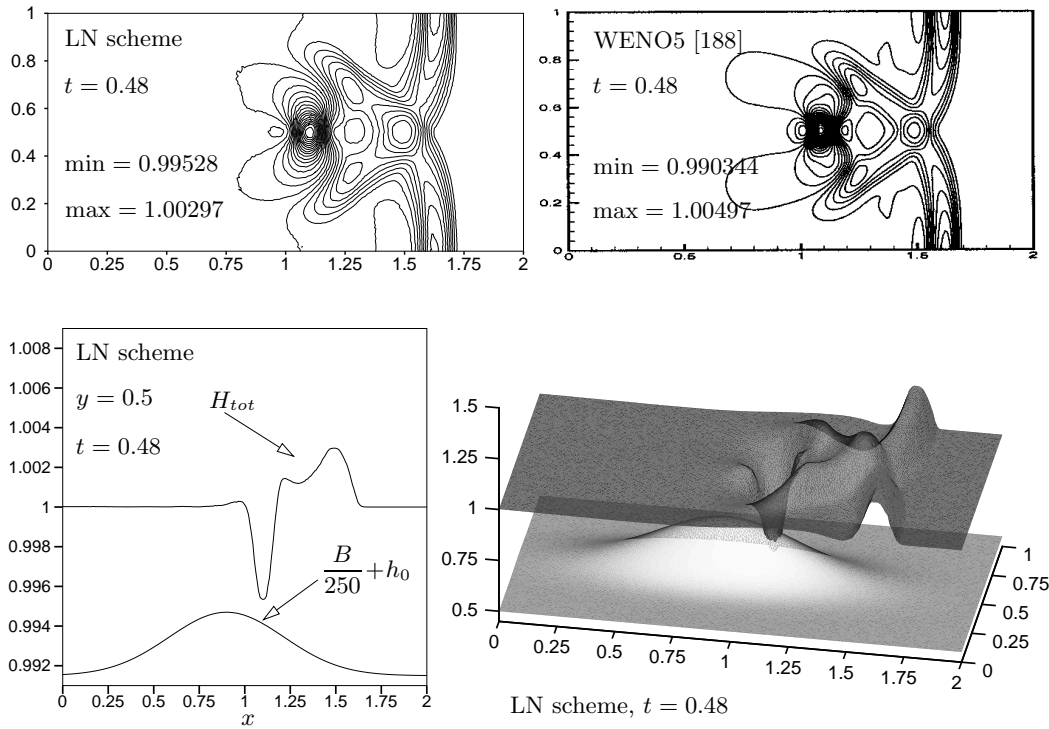


Figure 12.16: Water height perturbation over smooth bed. Solution of the LN scheme at time $t = 0.48$. Top-left: contour plot of total water height. Bottom-left: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$). Bottom-right: 3D plot of the solution. Top-right: contour plot of total water height from [188].

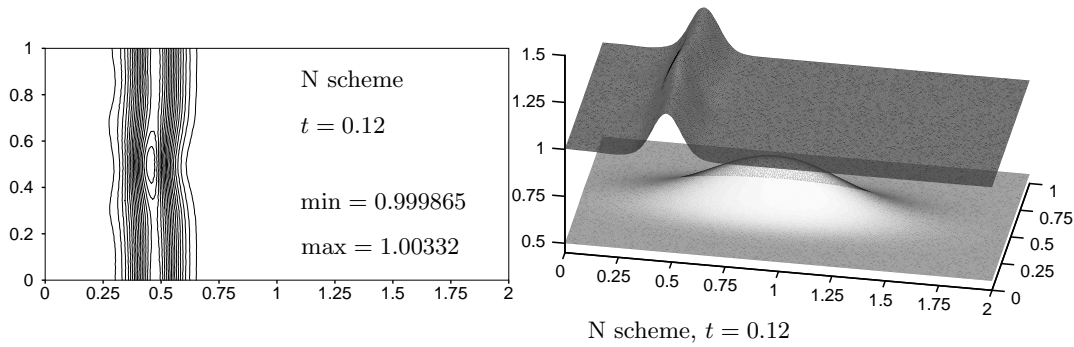


Figure 12.17: Water height perturbation over smooth bed. N scheme. Contours of total water height at time $t = 0.12$ (left) and 3D plot of the solution (right)

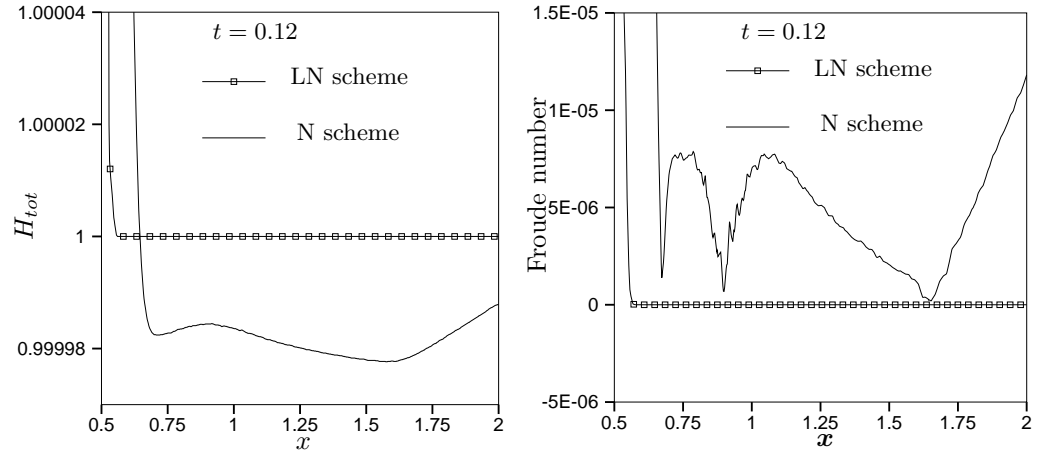


Figure 12.18: Water height perturbation over smooth bed. Total water height (left) and Froude number (right) in the unperturbed region (line $y = 0.5$) at time $t = 0.12$. N scheme (line) and limited N scheme (line with symbols).

12.3.4 Water height perturbation over non-smooth bed

We consider a variant of the previous problem involving a non-smooth variation of the bed height. In particular, we set

$$B(x, y) = 0.6e^{-\psi(x, y)}$$

with

$$\psi(x, y) = \begin{cases} \sqrt{(x - 0.9)^2 + (y - 0.5)^2} & \text{if } 0.3 \leq y \leq 0.7 \text{ and } 0.9 \leq x \leq 1.1 \\ 5(x - 0.9)^2 + 50(y - 0.5)^2 & \text{otherwise} \end{cases}$$

A 3D view of the scaled bed shape B^* (equation (12.9)) is reported in figure 12.19.

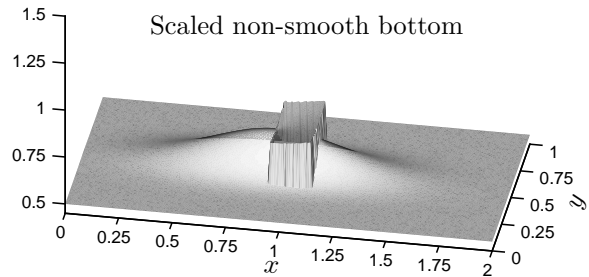


Figure 12.19: 3D view of the scaled non-smooth bed

12.3.4. Water height perturbation over non-smooth bed

The computational set-up and the initial state are identical to the ones used in §12.3.3. Similarly, the solution is qualitatively very close to the one obtained on the previous problem, until the perturbation reaches the discontinuity in $B(x, y)$. In particular, we report on figures 12.20, 12.21 and 12.22 the solution obtained with the LN scheme at times $t = 0.15$, $t = 0.30$ and $t = 0.45$. As before, in all the 3D plots (bottom pictures) we report on the third axis the scaled water and bed heights given by (12.9). Conversely, the scaling of the bed height used in the line plots on the top-right is indicated in the pictures. The remarks made for the previous test apply also to these results. The lake-at-rest solution is preserved exactly in the unperturbed region, despite of the non-smoothness of the shape of the bottom, and of the irregular mesh. Similarly, the total water height behind the perturbation gets back to a constant value very close to one, as clearly visible in figures 12.20 and 12.21.

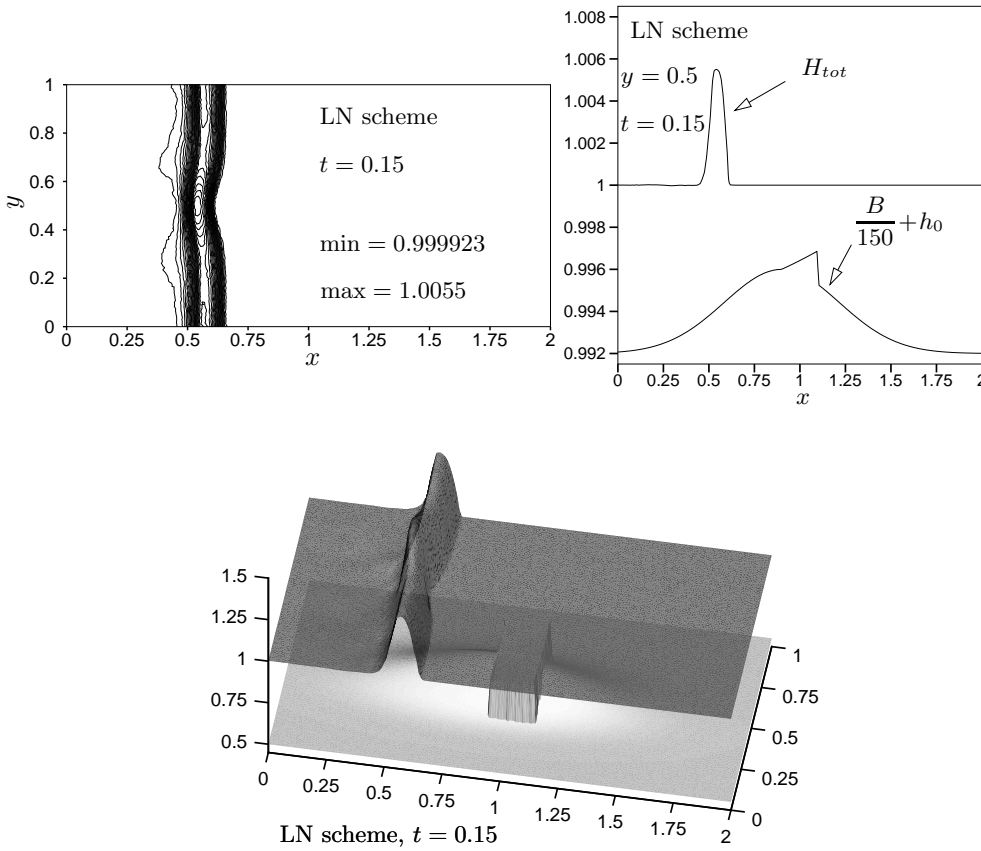


Figure 12.20: Water height perturbation over non-smooth bed. Solution of the LN scheme at time $t = 0.15$. Top-left: contour plot of total water height. Top-right: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.992$). Bottom: 3D plot of the solution.

The numerical solution obtained with the nonlinear scheme is quite stable and monotone, even if the data of the problem are non-smooth. Very small oscillations are present at later times of the simulation only in correspondence of the singular corners of $B(x, y)$, at $(x, y) = (0.9, 0.3)$, $(x, y) = (1.1, 0.3)$, $(x, y) = (0.9, 0.7)$ and $(x, y) = (1.1, 0.7)$. This is visible in figure 12.22. Due to the extremely low velocity and to the almost flat profile of H_{tot} , these oscillations are not dissipated by the scheme. As done for the previous problem, in figure 12.23 we compare the solution of the space-time N scheme with the one of LN scheme at time $t = 0.12$ in the unperturbed region, along the line $y = 0.5$: the LN scheme preserves the lake-at-rest state up to machine accuracy while very small perturbations are introduced by the first-order linear scheme.

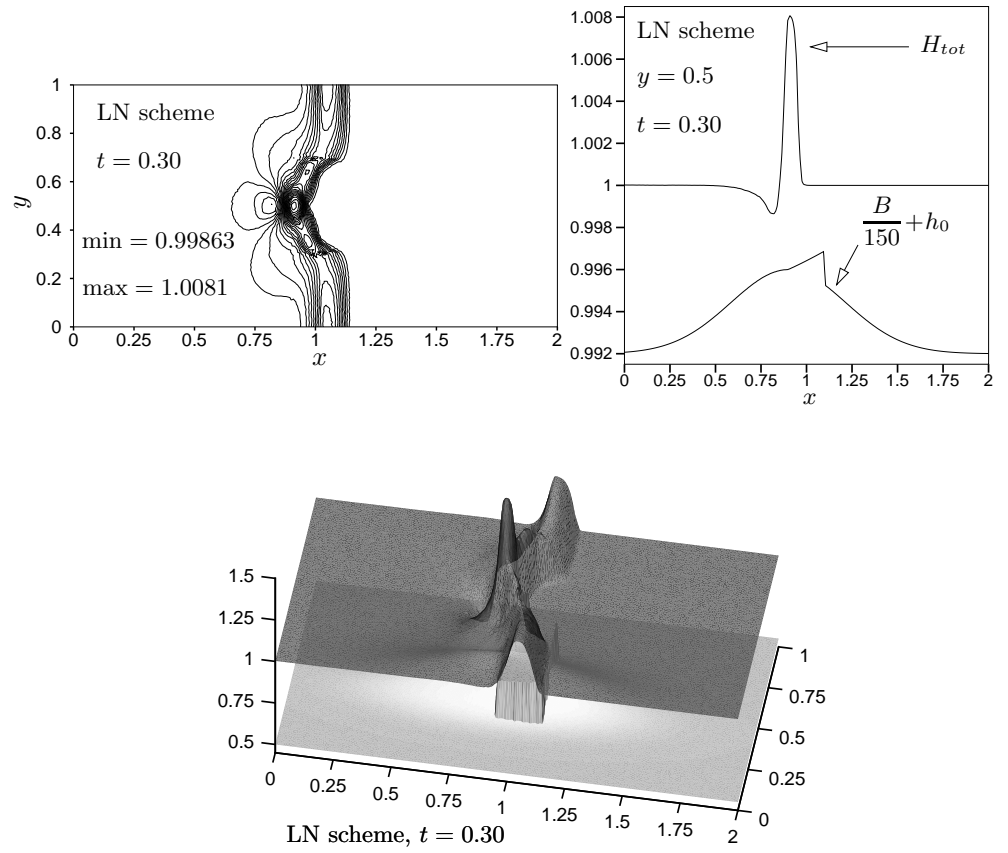


Figure 12.21: Water height perturbation over non-smooth bed. Solution of the LN scheme at time $t = 0.30$. Top-left: contour plot of total water height. Top-right: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.992$). Bottom: 3D plot of the solution.

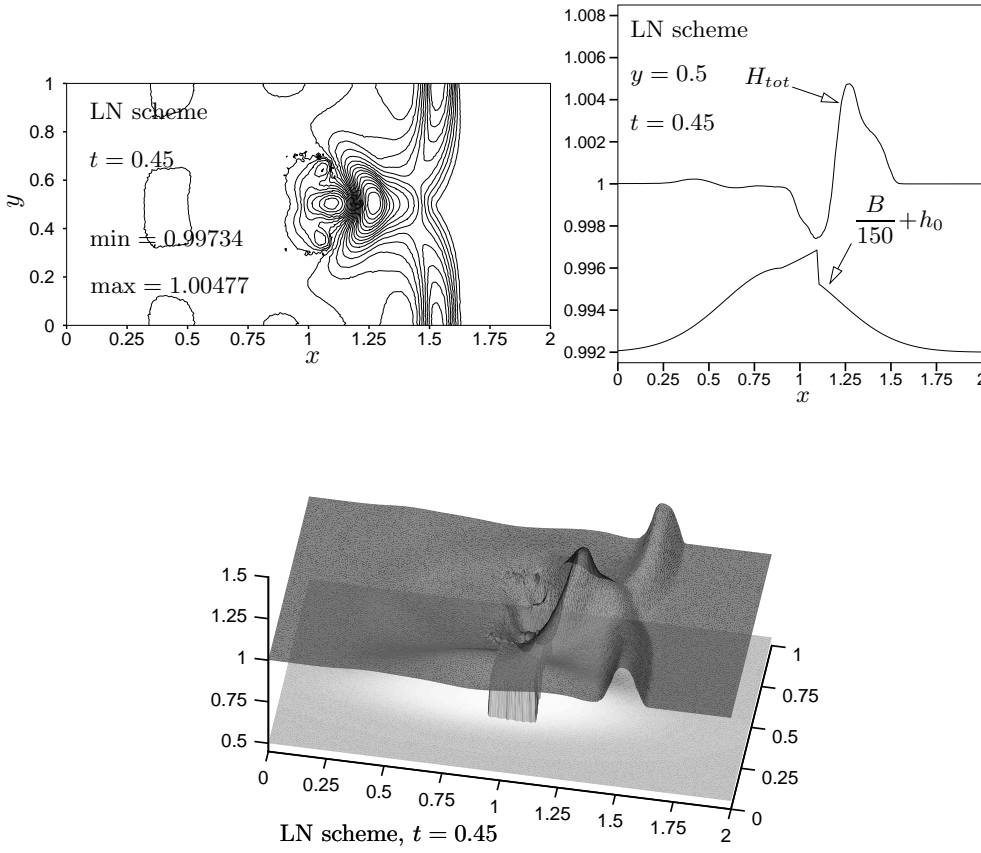


Figure 12.22: Water height perturbation over non-smooth bed. Solution of the LN scheme at time $t = 0.45$. Top-left: contour plot of total water height. Top-right: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.992$). Bottom: 3D plot of the solution.

12.4 Summary

In this chapter we have shown the application of our conservative schemes to the solution of the shallow-water equations. We have shown that the \mathcal{RD} framework allows easily to construct schemes which preserve *exactly* the lake-at-rest solution independently on the topology of the mesh, on the complexity of the bed shape and on the order of interpolation of the unknowns. We discussed the extensive numerical validation on shallow-water flows of the conservative formulations of the N scheme and of its limited variants on a wide number of steady and time-dependent problems involving flat and non-flat bed. The results show that indeed these schemes have a great potential for the computations of steady and time-dependent free surface shallow flows.

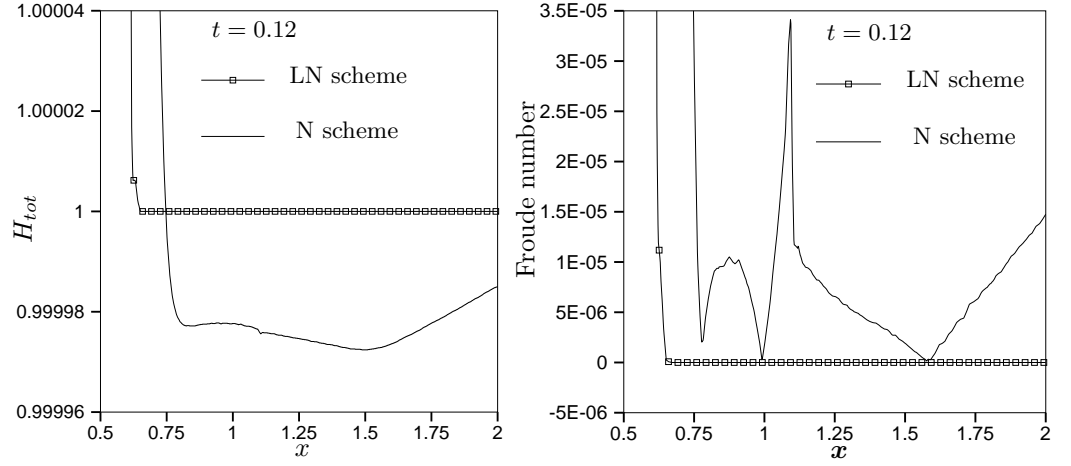


Figure 12.23: Water height perturbation over non-smooth bed. Total water height (left) and Froude number (right) in the unperturbed region (line $y = 0.5$) at time $t = 0.12$. N scheme (line) and limited N scheme (line with symbols).

However, we have left open one of the most challenging issues related to this topic: the computation of flows over dry bed. Due to the extensive use of the eigenstructure of the flux Jacobians, the singularities introduced by vanishing relative water height cannot be handled by the schemes proposed here. Dry shallow-water flows are equivalent to flows containing *vacuum* for the schemes. As for Roe's scheme, this state cannot be handled in a stable manner by our upwind \mathcal{RD} distribution. A possible solution to this problem could be the use of distribution strategies in which the dissipation does not depend heavily on the eigenstructure of the Jacobians, as for example in the Rusanov scheme of §6.2.2.3 and §6.3.1.3. However, this will be possible only after gaining a better understanding of the limiting technique used to generate the nonlinear schemes.

Chapter 13

Conclusions and perspectives

This thesis has presented the construction, the analysis and the verification of compact residual discretizations for the solution of conservation laws on unstructured meshes. The schemes considered belong to the residual distribution (\mathcal{RD})/fluctuation splitting (\mathcal{FS}) class. The design methodology presented relies on three main elements

1. Construction of compact, linear, first-order, monotone and stable schemes for linear hyperbolic PDEs;
2. A *positivity preserving* procedure mapping stable first-order linear schemes onto nonlinear second-order schemes with non-oscillatory shock capturing capabilities;
3. A conservative formulation enabling to extend the schemes to nonlinear \mathcal{CL} s.

These three *design* steps, and the underlying theoretical tools, have been discussed in depth. The nonlinear \mathcal{RD} schemes resulting from this construction have been tested on a large set of problems involving the solution of scalar models, and systems of \mathcal{CL} s. This extensive verification fills the gaps left open, where no theoretical results can be shown. On irregular grids, the schemes proposed yield quite accurate and stable results even on very difficult computations. These results are more accurate than the ones given by \mathcal{FV} and WENO schemes. Moreover, our schemes have a compact nearest-neighbor stencil. This encourages to further develop our approach, toward the design of very high-order schemes, which would represent a very appealing alternative, both in terms of accuracy and efficiency, to now classical \mathcal{FV} and ENO/WENO discretizations. These schemes might also be very competitive with respect to very high-order \mathcal{DG} schemes.

In this chapter we discuss in detail the main contributions of this work, and the performances of the schemes proposed. An attempt is made to isolate the elements rendering the technology used here better than other approaches, as well as the weak points of the construction. In this light, we will thoroughly describe the future perspectives of our discretization philosophy.

13.1 Main achievements

The thesis describes the construction of schemes for the solution of systems of \mathcal{CL} s. The three design phases mentioned before have been gradually developed, starting from the analysis of schemes for scalar advection, and arriving to nonlinear systems at the end of the manuscript. Most of the theoretical results, from the definition of monotonicity and stability, to the derivation of conditions allowing to achieve second-order of accuracy, have been discussed for scalar advection, and then applied to the analysis of scalar \mathcal{FS} discretizations. The schemes obtained in this way have been extended to the solution of nonlinear \mathcal{CL} s by introducing a quite general conservative framework. A *matrix* generalization allows to handle nonlinear systems. This has permitted to perform an extensive validation of our approach. All these elements have led to new developments, or have required the application of known analytical tools, still yielding results largely unpublished and allowing to have a *fresh look* on \mathcal{RD} schemes.

13.1.1 Compact cell-vertex schemes for scalar advection

All the elements needed to obtain stable and accurate discretizations, namely the definition of discrete monotonicity, stability and accuracy conditions, have been discussed for scalar advection. Even though the steady-state and time-dependent cases are considered separately in the manuscript, the same theoretical tools are used in the analysis.

The definition of discrete monotonicity has been achieved by making use, in the cell-vertex framework of the thesis, of the theory of positive coefficients [161]. This has ultimately led to conditions for the satisfaction of a discrete maximum principle. In the simplest case of explicit forward Euler time integration, the analysis presented here is a re-adaptation of the theory discussed in [20, 25] for \mathcal{FV} schemes on the median dual cell. However, the design of monotone space-time schemes for time-dependent calculations has required the extension of the analysis to implicit two-level time discretizations. Further generality has been obtained by considering non-homogeneous problems in which the source term does not depend on the solution. The analysis shows that even implicit schemes are subject to a time-step constraint for the preservation of the monotonicity. This is in line with the results obtained in [27] by studying ODEs representing the temporal evolution of monotone operators.

To the L^∞ stability associated to the maximum principle, we have added the L^2 stability related to the satisfaction of bounds on the energy of the discrete solution. This condition characterizes the dissipation present in the discretization. A quite general framework has been presented, showing that schemes with a Local Extremum Diminishing (LED) character introduce a degree of dissipation, modulo terms related to the boundary conditions. How to include these terms has been shown on some examples of central schemes. The analysis has been extended to the fully discrete case, introducing the concept of energy dissipation in time. For the implicit two-level time discretizations considered here, this additional stabilizing contribution becomes larger as the degree of implicitness of the time discretization increases.

For cell-vertex schemes based on a continuous representation of the unknowns on unstructured meshes, we have recalled and extended the conditions allowing to achieve second-order of accuracy. Following [9, 3, 12], this has been achieved by manipulating the discrete equations in a way that allows to recast them as a discrete variant of the definition of a weak-solution, plus some *scheme-dependent* extra terms, characterizing the consistency and the order of the approximation. This technique has been used to derive accuracy conditions in the homogeneous and non-homogeneous case. It has also allowed to give some formal evidence of the fact that, during transients, *mass-lumped* schemes cannot be second-order accurate on unstructured grids and, lastly, to analyze the accuracy of the space-time discretizations used in the thesis.

13.1.2 Fluctuation splitting schemes

The \mathcal{FS} schemes object of the thesis have been discussed and analyzed in considerable detail. For linear scalar advection, the theoretical analysis of cell-vertex discretizations performed in the thesis has given formal tools to precisely characterize their properties. Both in the steady and in the time-dependent case, analogies with \mathcal{FE} schemes have been used to illustrate some of these properties. However, \mathcal{RD} schemes have features of their own, which can hardly be incorporated in different discrete frameworks.

One of these features is the residual character of second-order schemes. In steady and unsteady computations, in the homogeneous and non-homogeneous case, the \mathcal{RD} approach allows the construction of true residual discretizations. These schemes are obtained by defining local nodal residuals proportional, through some uniformly bounded constant, to a local approximation of the integral of the equations. By construction, these *linearity preserving* (\mathcal{LP}) schemes meet the formal accuracy conditions proved in the general case. With respect to \mathcal{FV} and also \mathcal{DG} discretizations, the advantage is that the residual character of these schemes is independent on the stabilization procedure. It is intuitively and formally very clear to understand.

The stability of \mathcal{RD} schemes also relies on a mechanism peculiar to this framework: the *multidimensional upwinding* (\mathcal{MU}). The most successful \mathcal{MU} schemes, the LDA and the N schemes, have been analyzed in some detail. The LDA scheme is \mathcal{LP} . Hence its accuracy is guaranteed by its residual character, as confirmed by the scalar numerical experiments discussed in the thesis. Being linear, the LDA scheme cannot be also monotone, as stated by Godunov's theorem. However, its energy stability can be studied. This analysis has revealed the beneficial dissipative mechanism of the \mathcal{MU} . This unconventional stabilization has been shown to have a true multidimensional character. Even though the analysis has not led to real stability estimates, the analogy with SUPG schemes has revealed that the underlying mechanism is the same used in \mathcal{FE} : the addition of an anisotropic dissipation to a centered discretization. However, for \mathcal{MU} schemes this central discretization acts on streamlines, or characteristic lines. Conversely, the N scheme has very clear stability properties, while being only first-order accurate. Not only it enjoys all the monotonicity conditions proved in the thesis, but it also has a clear dissipative character. In particular, after recalling the energy stability

analysis of the \mathcal{N} scheme, we have shown that it can be written as the LDA scheme plus an anisotropic dissipation term. In the non-homogeneous case, this formulation reduces to the variant of the \mathcal{N} scheme proposed in [151]. When the source term does not depend on the solution, this variant is L^∞ -stable.

The construction of nonlinear \mathcal{RD} schemes for scalar advection has been discussed in depth. We have recalled the blending approach, showing its equivalence with the introduction of nonlinear shock-capturing dissipation. Also, we have presented a geometrical construction of the blended LDA/ \mathcal{N} scheme of [3]. However, the most appealing approach to obtain nonlinear \mathcal{FS} schemes is the limiting of a linear positive first-order scheme [129, 126, 9, 10, 12]. Again, this is a technique which is strongly tied to the \mathcal{RD} character of the discretization. It combines the residual property of \mathcal{LP} schemes with a positivity preserving mapping procedure which allows to generate monotone second-order discretizations starting from linear first-order positive schemes. In this thesis we have given formal conditions for the well-posedness of the limiting, showing the importance of the consistency of the linear positive scheme used in the mapping. As already remarked, this technique has a very strong L^∞ flavor, as far as the stability of the resulting nonlinear schemes is concerned. This is in contrast with what is done in the \mathcal{FE} context, where the global L^∞ stability of the discrete solution is a result of the regularization introduced by large local dissipation terms [166]. Hence, we could say that the \mathcal{FE} approach has a marked L^2 flavor. As a result, the dissipation properties of the limited nonlinear \mathcal{RD} schemes are not very clear. Here, we have discussed this issue in some detail, after [18]. The general conclusion is that this technology works best if the discretization has a marked upwind character, which, even though not rigorously proved, stabilizes the scheme. This is confirmed by the numerical experience. Conversely, the monotonicity and shock capturing properties of these schemes are excellent. Their accuracy on irregular meshes is confirmed by the numerical tests.

Concerning the time-dependent case, this thesis clearly shows that second-order \mathcal{RD} schemes need the introduction of a mass-matrix. We have also given formal evidence that, on general triangulations, the explicit Lax-Wendroff scheme proposed in [90, 60] cannot be second-order accurate. The space-time framework introduced in [8, 51, 120], and used in this work, is the only one satisfying all the design requirements for second-order schemes, while encompassing a large family of monotone schemes and allowing to use all the technology developed for steady-state calculations, including the construction of nonlinear high-order monotone schemes.

13.1.3 Residual distribution for nonlinear conservation laws

The solution of nonlinear conservation laws requires a conservative formulation of the method. In most of the literature this issue is handled by introducing local exact mean-value linearizations of the quasi-linear form of the problem. This allows to immediately apply, on the linearized problem, the schemes developed for scalar advection. The main limitation of this procedure is that the conservative linearization is only available for simple \mathcal{CL} s and on triangular elements. A more general approach, however based on the

same principles, is the one introduced in [4]. As recalled in the thesis, this technique aims at replacing exact mean-value flux Jacobians with approximate ones, obtained by Gaussian integration. If the number of quadrature points is large enough, the error introduced by the inexact evaluation of the mean-value flux Jacobians is within the discretization error, thus conservation is still guaranteed. This technique is general enough to allow the extension of the schemes to arbitrary \mathcal{CL} s and meshes, and it makes easy their formal analysis. However it is quite expensive due to the surface (volume in 3D) Gaussian quadrature. This is even more relevant in view of the extension to systems of \mathcal{CL} s for which the Jacobians are matrix functions.

The approach used in this thesis is more efficient than the one of [4]. To our knowledge, it represents the best solution available at the moment to extend \mathcal{FS} schemes to the discretization of very general \mathcal{CL} s and to meshes containing non-triangular elements. As originally proposed in [50] it is based on two main ingredients. The first is the direct use of the integral form of the \mathcal{CL} to define the local residual. This alone guarantees that a discrete analog of the Rankine-Hugoniot jump conditions is satisfied. The second ingredient is a re-formulation of the N scheme, guaranteeing its consistency with the conservative definition of the residual. In [50] this variant of the N scheme has been experimentally proved to be extremely robust and to yield non-oscillatory results. The definition of the cell residual via the contour integration of the fluxes on the boundary of the element has led the authors of [50] to refer to this approach as to the \mathcal{CRD} approach. These \mathcal{CRD} schemes guarantee discrete conservation independently on the linearization used to evaluate the flux Jacobian needed for the \mathcal{MU} procedure. The formulation introduced in [50] has later been used in [44, 140, 134, 63, 141, 142] to construct conservative formulations for the Magneto-Hydrodynamics equations, for time-dependent problems, and to design conservative schemes on grids composed of quadrilaterals. Even though part of the work contained in the manuscript was already published in [141, 142], with respect to these references the thesis gives a more mature description of this approach, and it provides some extra understanding of its properties.

By definition, a conservative \mathcal{RD} scheme is one for which the element residual is exactly equal to the contour integral of the fluxes on the boundary of the element, for some continuous discrete approximation of the flux. While, as shown in the thesis, this definition encompasses the schemes based on the exact mean-value linearization of the Jacobians, it obviously implies that the direct approximation of the contour integral in the definition of the residual yields conservative schemes. As far as the definition of the \mathcal{CRD} N scheme is concerned, we have shown that the formulation of [50] is a natural consequence of the fact that the N scheme can be written as the LDA plus dissipation terms. It is trivially obtained by keeping unchanged the analytical form of these extra terms, while still using the LDA scheme to distribute the contour integral of the fluxes. This perspective allows to give a heuristic justification of the fact that the behavior observed in practice for the \mathcal{CRD} N scheme is almost identical to the one of the scheme based on the conservative linearization: the underlying dissipation mechanism is the same. The \mathcal{CRD} formulation, completely decouples the issues of conservation and of the Jacobian linearization needed for the multidimensional upwinding, allowing to use arbitrary linearizations for this purpose. The effect of the use of inexact linearizations in the distribution of the residual has been analyzed. For the \mathcal{CRD} N scheme, the most

important is the loss of formal positivity, especially in multi-target elements. However, the non-positive coefficients eventually present in the discretization are proportional to the difference between the exact mean-value Jacobians and the inexact ones. Hence, this effect is globally very weak, as confirmed by the robustness and monotonicity that the scheme shows experimentally. The application of the limiting technique to the \mathcal{CRD} N scheme leads to a nonlinear \mathcal{LP} discretization showing excellent shock capturing and high-accuracy in smooth regions.

A second effect of the use of inexact linearizations in the distribution is related to the entropy stability of the schemes. As the energy stability in the linear case, in the nonlinear case the entropy stability of a scheme characterizes its dissipative behavior. This thesis gives some insight into the entropy dissipation properties of \mathcal{RD} schemes. As in the linear case, this analysis is supported by the comparison with \mathcal{FE} discretizations, for which stability estimates can be derived very easily. Unfortunately, this is not true for \mathcal{RD} discretizations, for which the weaker concepts of entropy consistent and entropy dissipative schemes have been introduced. Entropy dissipative schemes are the ones respecting an entropy inequality in the limit $h \rightarrow 0$. As in the linear case, we have been able to show that the multidimensional upwinding is very beneficial in terms of entropy stabilization. The dissipation mechanism is analog to the one acting in the linear case. The analysis of some \mathcal{CRD} linearity preserving schemes shows that the use of an inexact Jacobian linearization for the distribution leads to a loss of information related to the entropy dissipative character of the discretization, at least in multi-target elements. Concerning the N scheme, we have recalled the proof of [4], showing the entropy dissipative character of the scheme if exact mean-value Jacobians are used. Then, we have discussed the stability of the \mathcal{CRD} N scheme. We have shown the presence of an entropy dissipation mechanism, however, details are missing (or not entirely understood) to obtain a formal proof that the scheme is entropy dissipative. The analysis has been extended to the fully discrete time-dependent case, after [171]. As in the linear case, implicit schemes have the highest degree of entropy dissipation, while explicit schemes add entropy *destabilizing* terms, which need to be balanced by the dissipation of the spatial discretization.

Still concerning the time-dependent case, we have presented a consistent extension of the \mathcal{CRD} formulation to the space-time framework used in the thesis, following the initial work reported in [141, 142]. In particular, the combination of \mathcal{CRD} space-time variants of the N scheme with the limiting technique leads to nonlinear \mathcal{LP} conservative schemes for arbitrary time-dependent conservation laws on unstructured meshes. Numerical results show excellent shock capturing and high-accuracy in smooth regions.

13.1.4 Systems of \mathcal{CL} s: verification

In this thesis the extension to systems is obtained in a formal way by resorting to the *matrix* formulation of \mathcal{RD} introduced in [177, 178]. This approach is quite straightforward, however, it leads to the loss of all the geometrical analogies allowing to analyze the scalar schemes. Even so, most of the properties, such as linearity preservation and

multidimensional upwinding, can be generalized to the matrix schemes. The energy and entropy stability analyses also admit a formal generalization. The stability properties of matrix \mathcal{MU} schemes are the same of their scalar counterparts, the upwinding still introducing dissipation. Conversely the definition of a monotone scheme is more difficult, even though the L^∞ analysis on simple waves of [10, 9, 118] can be used.

The matrix formulation of the \mathcal{CRD} N scheme, of its space-time variants, and of the nonlinear conservative \mathcal{LP} matrix schemes obtained by limiting these linear schemes, has allowed to perform a very extensive verification of our methodology on the solution of several systems of \mathcal{CL} s. The aim of these experiments has been to verify, in different settings, the robustness, the monotonicity and the ability of the schemes proposed to resolve complex flow structures on irregular grids. All the tests performed have a somewhat academic character. They involve the solution of the Euler equations of a perfect gas, of a two-phase flow model, and of the shallow-water equations.

The literature is full of test problems and reference results based on the solution of the Euler equations of a perfect gas. This motivates the continuous use of some of these tests for the verification of new discretization techniques. On these equations the schemes proposed in this thesis have performed extremely well. The monotonicity of the \mathcal{CRD} N scheme and of its space-time variants has been largely confirmed. More importantly, the nonlinear limited schemes have proved very robust even in presence of strong steady and unsteady discontinuities. Concerning the resolution of complex structures, the comparison with the \mathcal{FV} scheme of [24] has shown that the technology used in this thesis definitely leads to more accurate discretizations on unstructured meshes. Grid refinement studies have also shown the ability of the schemes proposed to compute unsteady inviscid instabilities without dissipating them. These results are analog to the ones obtained with second-order \mathcal{DG} discretizations.

The two-phase flow equations considered in the thesis constitute a simple model of homogeneous air/water flow. However, the nonlinearity of the algebraic relations defining the underlying thermodynamics is such that no conservative mean-value linearization of the flux Jacobians can be found. Besides, the pressure cannot be expressed in closed form as a function of the conserved quantities. Hence, this model has all the features of systems of conservation laws with complex thermodynamics, for which the use of the conservative framework of the thesis is necessary. Not only the conservative character of the \mathcal{CRD} discretization has been confirmed by the numerical results, but the excellent shock capturing obtained with the limiting approach, and the high resolution of the nonlinear schemes have also been proved. This in a more complex setting, involving stronger nonlinearities.

The solution of the shallow-water equations on non-flat bed is an application of considerable engineering interest. In two space dimensions this system of equations does not admit a simple conservative mean-value linearization, making the \mathcal{CRD} formulation proposed here well suited. The simulations performed involve frictionless flows without dry areas. The results on flat bed topologies confirm the conservative character of the schemes, their monotonicity and robustness, and the high resolution of the limited variants of the N scheme. Simulations of flows on non-flat bed involve the discretization of

the source term modeling the variation of the bed height. When writing the equations in terms of local water height and velocity, this source term is independent on the solution. This justifies the application of the matrix variant of the schemes developed for non-homogeneous scalar advection. The numerical experiments have confirmed their monotonicity. Moreover, on this application the residual character of second-order \mathcal{RD} schemes has a very interesting consequence: steady-state lake-at-rest solutions of the shallow-water equations are preserved exactly by \mathcal{LP} schemes. This theoretical result, confirmed by the numerical simulations, is a quite natural consequence of the residual approach used here, in contrast with the more complex constructions needed in the \mathcal{FV} context to obtain discretizations with similar properties. Besides, this property of the \mathcal{LP} schemes can be proved for more general exact solutions [138], giving additional evidence of the impressive potential of the residual-based approach.

13.2 Weaknesses of the methodology proposed

The construction presented in this thesis leads to schemes with a strong non-oscillatory character. These schemes also enjoy a residual property formally guaranteeing second-order of accuracy. The numerical results show a very good potential in terms of accuracy and robustness. However, the approach used to construct these schemes has some limitations, which we will underline in the following. These *weaknesses* are related to the basic issue of stability of the limited nonlinear schemes and of the computational cost of the discretization. For high-speed or wave-propagation problems the matrix upwind technology used here pays off in terms of robustness, accuracy and stability. This is especially true for the nonlinear limited schemes. Things are different for low-speed and transonic problems, for which the strongly non-oscillatory character of the discretization is not as important, and where stabilized central discretizations perform quite well, rendering the use of an expensive upwinding procedure unjustified. There is room and need for improvements.

13.2.1 Nonlinear schemes and stability

As remarked more than once, the mapping of a positive linear first-order scheme onto a nonlinear \mathcal{LP} scheme leads to a monotone discretization, provided that positivity is preserved by the nonlinear mapping. In contrast to the standard shock capturing techniques used *e.g.* in \mathcal{FE} methods, this approach strongly enforces the L^∞ -stable character of the nonlinear scheme. The numerical results show that this technique is indeed very well suited for the design of monotonicity preserving schemes. However, in general the question of the L^2 stability remains open. As the grid convergence study of §7.3.1 shows, sometimes the accuracy measured for the schemes is below the expectations, despite of their \mathcal{LP} character. By construction, these schemes are consistent, hence lack of convergence could be related to the presence of a weak instability. Unfortunately, we are not able to formally justify this behavior, even though the analysis of §5.5.2.3 hints at the presence of an (energy) destabilizing mechanism.

Some numerical results in [10, 9, 118] show that, if applied to non-upwind linear positive schemes, the limiting technique leads to nonlinear discretizations performing very poorly. In the development of this thesis, we have been led to similar conclusions when trying to construct nonlinear discretizations for the advection-reaction equation, and limiting an \mathcal{N} scheme with the reaction term treated in a pointwise manner [144, 173]. The characteristic symptoms of the problem are a poor iterative convergence and a general lack of smoothness of the solution, as if an overcompressive mechanism was active. These symptoms are scarcely visible in the results of the thesis, even though the poor convergence of the nonlinear matrix schemes is undeniable. In the case of the advection-reaction equation, we have experimentally noted that the addition of a Least Squares-type stabilization term to the limited scheme cures both the iterative convergence and the smoothness problem, however spoiling the positivity of the discretization. This suggests that the origin of the problem is indeed related to a lack of dissipation. The absence of the problem when limiting linear first-order schemes with a marked upwind character, could be explained by the fact that the resulting limited scheme has itself a strong upwind bias. This has been seen to have a stabilizing effect. We also recall that in [12] the \mathcal{MU} has been shown to lead, in two space dimensions, to \mathcal{FS} discretizations with a stable character, in the sense of the coercivity of the bilinear form obtained by recasting the schemes in a variational form (see [12, 9] for more).

The understanding of this problem, as well as the design of a cure for it, is of paramount importance for the success of this technology. We believe that if the excellent shock capturing obtained with the limiting technique could be incorporated into an L^2 stable \mathcal{LP} discretization, \mathcal{RD} would have very high chances of success in the competition with other approaches, as for example \mathcal{DQ} . This as a consequence of the fact that, even for a second-order approximation, the gain in accuracy due to the recovery of the full convergence rate, and in efficiency, due to the possibility of using simpler linear first-order schemes in the mapping, coupled with the compact and true monotonicity preserving character of the schemes, would give unique stability and accuracy properties.

13.2.2 On the efficiency

A second, though not less important, problem is related to the computational cost of the schemes. For systems, this issue is strictly tied to the matrix formulation adopted. The use of multidimensional upwind matrix schemes leads to the need of numerically inverting on each element small full matrices. This operation is quite expensive and is probably where most of the computational time is lost. Even though, with respect to second-order \mathcal{FV} schemes, this is compensated by the fact that no multidimensional polynomial reconstructions are needed, there is room for improvement. This might be particularly important in view of the design of very high-order schemes for systems.

One solution could be to re-engage the study of scalar decompositions of the system of equations. For steady computations involving the solution of the Euler equations, this has led to quite successful results [126, 130, 123, 136]. In three space-dimensions and in the time-dependent case the degree of decoupling which can be introduced is lower.

Nevertheless, the reduction of the size of the matrices to invert of even one unit would already bring some savings. This, for example, can be achieved by using for the spatial distribution the quasi-linear form of the equations in symmetrizing variables, in which the entropy equation is always uncoupled [86, 87, 28]. This would still guarantee the generality of the formulation.

A more appealing alternative can be thought of. If an improved limiting procedure was at hand, allowing to operate on Lax-Friederichs type first-order schemes such as the Rusanov scheme of §6.2.2.3, there would be no need at all to evaluate the complete eigenstructure of the Jacobians, and the cost of the schemes would be mainly related to the evaluation of the element residual. The latter could be minimized by using continuous polynomial flux approximations of the minimum possible degree needed to guarantee the required accuracy. This approach would also free the schemes by the singularities induced by the intensive use of the quasi-linear form. Once the number of matrix operations is reduced to the minimum, the schemes might be very competitive with \mathcal{DG} discretizations also in terms of computational cost.

Lastly, we remark that in time-dependent computations the two-layers variant of the space-time schemes should be always preferred, especially when dealing with large meshes, or in presence of locally highly refined areas in the grid.

13.3 Future perspectives

The discretization approach used in this thesis has shown a potential which certainly justifies its further development. This can be done on different fronts. First of all, we remark that the parallel implementation of the schemes, completely overlooked in this manuscript, is of course essential to benefit from their compactness. Preliminary results on this topic, obtained at the von Karman Institute for Fluid Dynamics by J. Dobeš, have shown the advantage of the space-time matrix schemes over implicit second-order \mathcal{FV} schemes, which are penalized by need of performing the linear reconstruction, and exchanging between processors the related information. More interesting developments can however be foreseen.

13.3.1 Very high-order schemes

The design of \mathcal{FS} schemes of order of accuracy higher than two is of primary importance. Preliminary constructions have been presented in [12, 9, 139]. Simple experiments on scalar advection show that very high-order schemes are more efficient than second-order ones, in the sense that the strong reduction of the error due to the higher accuracy compensates the larger number of operations needed to obtain the very high-order approximation [181]. A similar behavior is observed in \mathcal{DG} [42]. In parallel implementations, this effect should be more pronounced, due to the fact that the number of local operations performed in the very high-order case is larger than for

second-order schemes. This allows to foresee a better parallel efficiency for the very high-order schemes, leading to an additional gain. These trends encourage to pursue the development of higher-order accurate \mathcal{RD} , refining the initial results of [12, 9, 139]. The most important aspect is the study of very high-order well-posed and stable non-linear schemes. The need of improved constructions clearly emerges in [12, 9, 139], and is confirmed by the analysis performed in this thesis (§5.5.2.2).

As for second-order schemes, the key of success of higher-order \mathcal{RD} will be a clever combination of the shock capturing obtained with the limiting technique and of the residual character of \mathcal{LP} schemes into a stable discretization. Moreover, in the very high-order case the possibility of using Lax-Friederichs type linear schemes as a basis for the limiting becomes rather important, due to the need of compensating the extra cost related to the higher-order interpolation with a simplification in the distribution. For systems this will be of paramount importance in the competition with \mathcal{DG} .

13.3.2 Viscous terms and sources

The approximation of viscous terms and of source terms dependent on the solution are important aspects, missing in the construction of the thesis. In most of the \mathcal{RD} literature, the discretization of the diffusive part of the Navier-Stokes equations is performed by resorting to a pure Galerkin approach. Despite of the analogy with linear continuous \mathcal{FE} schemes recalled in the thesis, there is no evidence that a Galerkin discretization of the diffusive terms, coupled with the upwind \mathcal{RD} treatment of the hyperbolic part of the equations leads to a second order accurate scheme. Indeed, in the \mathcal{FE} case, second-order schemes for advective-diffusive systems incorporate an appropriate scaling of the streamline dissipation terms with a local *Peclet* (or *Reynolds*) number. This scaling is essential for the accuracy of the discretization [94, 96, 97].

As remarked in the introduction of this thesis, in the \mathcal{RD} literature the interaction between the discrete transport (hyperbolic) operator and the discrete viscous operator has scarcely been addressed in the past¹, as underlined in the recent work of [124]. In the last reference, in particular, it is suggested to rewrite the second-order PDE as a first-order system to be discretized with a \mathcal{RD} approach. The scaling with the Peclet number is taken into account in the distribution. This approach works quite well in the \mathcal{DG} case, however it is not very well suited for \mathcal{RD} , for two main reasons. Firstly, due to the continuous variable approximation used in \mathcal{FS} , it leads to an excessive overhead in the cost of the schemes, due to the increased number of variables (conserved variables plus viscous stresses). Moreover, it makes very hard the definition of a monotone scheme, even in the scalar case. This is due to the fact that we are faced with the solution of a system. In the development of this thesis, a preliminary study of this issue has led to the conclusion that a variational discretization of the second-order differential terms can be still used, provided that the \mathcal{RD} approximation is written as a Galerkin scheme plus some extra terms. These are then weighted by a parameter scaling as a cell Reynolds number, exactly as in SUPG \mathcal{FE} schemes. The results

¹probably with the unique exception of the *off-stream* work of [29]

show that second and third-order linear schemes can be indeed obtained in this way [181, 146], and that second-order monotone schemes for steady and time-dependent problems can be designed with the exact same technology discussed in the thesis [145]. In time-dependent computations, the monotonicity of these schemes is however still constrained by a strict time-step limitation of the explicit type. This is unacceptable in the viscous case for an implicit scheme. The construction of unconditionally monotone discretizations for time-dependent advective-diffusive systems is still an open issue.

Concerning the \mathcal{RD} discretization of source terms dependent on the solution, as already remarked, the preliminary theoretical investigation performed during the development of this thesis [144, 173]¹ has shown that the most $(L^\infty-)$ stable approach is obtained with a pointwise treatment. This is in line with the experimental experience gained through the years in the approximation of source terms in turbulence models [176], or arising in chemically reacting flows [58], and also in the context of the shallow-water equations [92]. Unfortunately, the systematic construction of nonlinear schemes for general non-homogeneous problems proves quite difficult at the moment, as a consequence of the limited knowledge of the stability of nonlinear limited \mathcal{RD} schemes. Additional theoretical work is still needed in this area.

13.3.3 Hybrid, adaptive and moving meshes

As shown by the work reported in the volume [22], the efficiency of schemes developed for computational fluid dynamics purposes can be increased dramatically if hybrid, adaptive and moving meshes can be handled. This due to the possibility of optimizing number and position of grid entities, with respect to some error monitor. Moreover, for problems involving moving boundaries, the use of moving grids is a must.

The use of hybrid structured/unstructured grids is in theory possible, thanks also to the \mathcal{CRD} formulation of the schemes [134, 63]. Moreover, in the framework of \mathcal{RD} schemes, the use of solution-dependent grid adaptation has been already investigated in the past [127, 126]. Successful developments will however depend on the possibility of deriving systematic error-based adaptation strategies. In this respect, clear variational formulations of \mathcal{RD} need to be developed. One of such formulations is discussed in [12]. Alternatively, in the second-order case, the analogy with linear stabilized \mathcal{FE} might be useful. The most challenging issue will be the handling of *hanging* nodes in the framework of the \mathcal{RD} continuous approximation.

We also mention the work of [64], in which the nonlinear space-time scheme of [8] has been generalized to moving and deforming grids. Successful applications to the computation of flows with moving boundaries are shown in the reference.

¹not included in the manuscript due to its preliminary character

13.3.4 Applications

The conservative framework of the \mathcal{CRD} schemes lends itself to a multitude of interesting applications. In [50], where this conservative approach was originally proposed, the application to the solution of the Magneto-Hydrodynamics (MHD) equations was already shown, underlying the need of a better discrete approximation of the magnetic field. Truly solenoidal discrete formulations of the MHD equations on unstructured grids have been developed and are being improved by Á. Csík [45].

Concerning the shallow-water equations, the theoretical and numerical results contained in the thesis give a basis for the use of our schemes for the simulation of flows of some practical interest. Additional results are presented in [138], yielding a more complete overview of the possibilities of our approach. However, simulations of real engineering relevance will require the discretization of additional source terms, *e.g.* modeling bed friction, and the possibility of handling dry areas. As far as the bed friction is concerned, the development of robust discretizations for nonlinear source terms is needed. The computation of flows with dry areas could benefit from the use of schemes in which the dissipation does not depend extensively on the eigenstructure of the flux Jacobians, as in the \mathcal{MU} schemes used here. In both cases, the additional study of the properties of nonlinear limited \mathcal{RD} discretizations will be beneficial.

Another very interesting and challenging application is the solution of more general two-phase flow models. Certainly, this will require the development of robust discretizations of source terms. However, some more basic issues are present which deserve particular attention, such as the computation of flows in which one of the phases *vanishes*, and the approximation of strong contact discontinuities, especially in presence of large variations of the transverse velocity across the discontinuity. The first issue is similar to the problem of approximating shallow free-surface flows with dry areas, and boils down to the design of schemes able to handle *vacuum* in a stable manner. The computation of strong contact discontinuities has a more general relevance. It is known that this issue is already very important in single phase flow simulations [2, 11]. Even though it has been shown in [50] that a steady mesh-aligned shear can be exactly preserved by the \mathcal{CRD} schemes, numerical experience shows that this is not true on irregular meshes, where no alignment is possible. In these cases, very strong contacts lead to a wrong approximation of the pressure field. Simple experiments can be performed to check that, if the kinetic energy across the contact is artificially imposed, the approximation of the pressure improves quite a lot. This shows that at the basis of the problem there is a phenomenon very similar to the one observed in \mathcal{FV} [11]: the conservative method is unable to predict correctly the transverse kinetic energy and, ultimately, the pressure. Not surprisingly, in 1D the fix proposed in [2] can be rewritten in a \mathcal{RD} formalism, requiring the space-time residual to be consistent with the proper pressure and kinetic energy equation. This observation can probably be generalized to multiple space dimensions and could lead to understand how \mathcal{RD} schemes should be modified to yield the correct result. The experience gained in this way should help to analyze multi-species or multi-phase models of increasing complexity. It is believed that, due to the nature of the schemes, a proper approximation of the residual will always be the basis of these developments.

Lastly, we mention the ongoing work at the von Karman Institute for Fluid Dynamics on the extension of the schemes to the approximation of high enthalpy chemically reacting flows, using the thermodynamics library developed in [114, 115, 116]. To our knowledge, the work of [58] constitutes the only past attempt to perform this type of computations with \mathcal{RD} . The conservative framework of the \mathcal{CRD} formulation will certainly help to simplify the formulation of the schemes with respect to the reference. A careful study of the discretization of the source terms related to the chemistry will be needed.

Bibliography

- [1] R. Abgrall. A genuinely multidimensional Riemann solver. Technical Report 1859, INRIA, 1993.
- [2] R. Abgrall. How to prevent pressure oscillations in multi-component flows : a quasi-conservative approach. *J. Comput. Phys.*, 125(1):150–160, 1996.
- [3] R. Abgrall. Toward the ultimate conservative scheme : Following the quest. *J. Comput. Phys.*, 167(2):277–315, 2001.
- [4] R. Abgrall and T.J. Barth. Residual distribution schemes for conservation laws via adaptive quadrature. *SIAM J. Sci. Comput.*, 24(3):732–769, 2002.
- [5] R. Abgrall and K. Mer. Un théorème de type Lax-Wendroff pour les schémas distributifs. Technical Report 98010, Mathématiques Appliquées de Bordeaux, 1998.
- [6] R. Abgrall, K. Mer, and B. Nkonga. A Lax-Wendroff type theorem for residual schemes. In M. Hafez and J.J. Chattot, editors, *Innovative methods for numerical solutions of partial differential equations*, pages 243–266. World Scientific, 2002.
- [7] R. Abgrall and M. Mezine. A consistent upwind residual scheme for scalar unsteady advection problems. October 2000. Conference AMIF 2000, organized by the European Science Foundation. Tuscany, Italy.
- [8] R. Abgrall and M. Mezine. Construction of second-order accurate monotone and stable residual distribution schemes for unsteady flow problems. *J. Comput. Phys.*, 188:16–55, 2003.
- [9] R. Abgrall and M. Mezine. Residual distribution schemes for steady problems. *VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [10] R. Abgrall and M. Mezine. 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, B. Nkonga, and R. Saurel. Efficient numerical approximation of compressible multi-material flow on unstructured meshes. *Computer and Fluids*, 32:571–605, 2003.

- [12] R. Abgrall and P.L. Roe. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.*, 19(3):3–36, 2003.
- [13] R.A. Adams. *Sobolev spaces*. Academic Press, 1978.
- [14] V.I. Arnold. *Lectures on partial differential equations*. Springer-Verlag, Heidelberg, 2000.
- [15] M. Arora and P.L. Roe. On post-shock oscillations due to shock capturing schemes in unsteady flows. *J. Comput. Phys.*, 130:25–40, 1997.
- [16] A. Athanasiadis. *Three-dimensional hybrid grid generation and application to Navier-Stokes computations*. PhD thesis, Université Libre de Bruxelles, 2004.
- [17] T.J. Barth. Aspects of unstructured grids and finite-volume solvers for the Euler and Navier-Stokes equations. in *VKI LS 1994-05, Computational Fluid Dynamics Course, von Karman Institute for Fluid Dynamics*, 1994.
- [18] T.J. Barth. An energy look at the N scheme. Working notes, NASA Ames research center, CA, USA, 1996.
- [19] 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.
- [20] T.J. Barth. Numerical methods for conservation laws on structured and unstructured meshes. *VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [21] 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*. Springer-Verlag, Heidelberg, 1999.
- [22] T.J. Barth and H. Deconinck, editors. *Error estimation and adaptive discretization methods in CFD*, volume 25 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Heidelberg, 2003.
- [23] T.J. Barth and P.O. Frederickson. High-order solution of the Euler equations on unstructured grids using quadratic reconstruction. AIAA paper 90-0013, January 1990. 28th AIAA Aerospace Sciences Meeting, Reno, Nevada (USA).
- [24] 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).
- [25] T.J. Barth and M. Ohlberger. Finite volume methods: foundation and analysis. In E. Stein, R. de Borst, and T.J.R. Hughes, editors, *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Ltd., 2004.
- [26] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

-
- [27] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discétization des problèmes d'évolution paraboliques. *R.A.I.R.O. Analyse Numérique*, 12:237–254, 1978.
- [28] A. Bonfiglioli and H. Deconinck. Multidimensional upwind schemes for the 3D Euler equations on unstructured tetrahedral meshes. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 141–185. Vieweg, Braunschweig, 1997.
- [29] 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.
- [30] S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods – Second edition*. Springer, 2002.
- [31] H. Brezis. *Analyse Fonctionnelle: Théorie et Applications*. Masson, Paris, 1984.
- [32] D. Caraeni, Ch. Bergström, and L. Fuchs. Parallel NAS3D: an efficient algorithm for engineering spray simulations. In *Proc. of ParCFD99, Williamsburg*. Elsevier, 2002.
- [33] D. Caraeni, M. Caraeni, and L. Fuchs. A parallel multidimensional upwind algorithm for LES. 15th AIAA Computational Fluid Dynamics Conference, Anaheim, CA, USA, June 2001.
- [34] D. Caraeni, S. Conway, and L. Fuchs. About a parallel multidimensional solver for LES. Finite Volumes for Complex Applications II. Hermes Science Publications, Paris, 1999.
- [35] D. Caraeni and L. Fuchs. LES using a parallel multidimensional upwind solver. ICCFD, First International Conference on Computational Fluid Dynamics, Kyoto, Japan, July 2000.
- [36] D. Caraeni and L. Fuchs. A new compact high-order multidimensional upwind discretization. 4th World CSCC conference, Vouliagmeni, Greece, July 2000.
- [37] D. Caraeni and L. Fuchs. Compact third-order multidimensional upwind scheme for Navier-Stokes simulations. *Theoretical and Computational Fluid Dynamics*, 15:373–401, 2002.
- [38] 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.
- [39] 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.

- [40] L.A. Catalano, P. De Palma, M. Napolitano, and P. Pascazio. Genuinely multi-dimensional upwind methods for accurate and efficient solutions of compressible flows. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 221–250. Vieweg, Braunschweig, 1997.
- [41] L.A. Catalano, P. De Palma, M. Napolitano, and P. Pascazio. A very-efficient local-adaptive multigrid method based on a simple-wave decomposition of the Euler equations. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 187–220. Vieweg, Braunschweig, 1997.
- [42] 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.
- [43] C. Corre. Personal communication.
- [44] Á. Csík. *Upwind Residual Distribution Schemes for General Hyperbolic Conservation Laws and Application to Ideal Magnetohydrodynamics*. PhD thesis, Katholieke Universiteit Leuven, Faculteit Wetenschappen Centrum voor Plasma-Astrofysica, Belgium, 2002.
- [45] Á. Csík. A solenoidal discretization of the MHD equations on unstructured grids using upwind conservative residual distribution schemes. Submitted to *J. Comput. Phys.*, 2005.
- [46] Á. Csík and H. Deconinck. Space time residual distribution schemes for hyperbolic conservation laws on unstructured linear finite elements. In M.J. Baines, editor, *ICFD Conference on Numerical Methods for Fluid Dynamics VII*, pages 557–564, Oxford, 2001.
- [47] 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.
- [48] 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.
- [49] 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.

-
- [50] 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.
- [51] 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, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [52] Á. Csík, M. Ricchiuto, and H. Deconinck. Space-time residual distribution schemes for two-dimensional Euler and two-phase flow simulations. In J. Piau, M. Champion, J.-J. Gagnepain, O. Pironneau, B. Stoufflet, and Ph. Thomas, editors, *Fluid Dynamics and Aeronautics New Challenges, A Series of Handbooks on Theory and Engineering Applications of Computational Methods*. CIMNE Barcelona, 2003.
- [53] 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.
- [54] H. Deconinck and G. Degrez. Multidimensional upwind residual distribution schemes and applications. *Finite Volumes for Complex Applications II*. Hermes Science Publications, Paris, 1999.
- [55] H. Deconinck, M. Ricchiuto, and K. Sermeus. Introduction to residual distribution schemes and stabilized finite elements. *VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [56] H. Deconinck, P.L. Roe, and R. Struijs. A multidimensional generalization of Roe’s difference splitter for the Euler equations. *Computer and Fluids*, 22(2/3):215–222, 1993.
- [57] 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).
- [58] 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.
- [59] A.I. Delis and Th. Katsaounis. Relaxation schemes for the shallow water equations. *International Journal for Numerical Methods in Fluids*, 41:695–719, 2003.
- [60] P. De Palma, G. Pascasio, and M. Napolitano. An accurate fluctuation splitting scheme for the unsteady two-dimensional Euler equations. ECCOMAS CFD Conference, 2001, Swansea, Wales, UK, September 2001.
- [61] P. De Palma, G. Pascasio, G. Rossiello, and M. Napolitano. Accurate solutions to unsteady problems by monotone implicit fluctuation splitting schemes. 16th AIAA Computational Fluid Dynamics Conference, Orlando, Florida, USA, 2003.

- [62] 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.
- [63] P. De Palma, G. Pascazio, D.T. Rubino, and M. Napolitano. Multidimensional upwind cell-vertex schemes for quadrilaterals. ECCOMAS CFD Conference 2004, Jyväskylä, July 2004.
- [64] 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).
- [65] J. Dobeš, M. Ricchiuto, and H. Deconinck. Implicit space-time residual distribution method for unsteady laminar viscous flow. *Computer and Fluids*, 34:617–640, 2005.
- [66] J. Donea and A. Huerta. *Finite Element methods for flow problems*. John Wiley and Sons, UK, 2003.
- [67] L. C. Evans. *Partial differential equations*. AMS Press, 1998.
- [68] L.C. Evans. Entropy and partial differential equations – Lecture notes. <http://math.berkeley.edu/~evans/entropy.and.PDE.pdf>.
- [69] R.P. Fedkiw, T. Aslam, B. Mettiman, and S. Osher. A non-oscillatory eulerian approach to interfaces in multi-material flows (the ghost fluid method). *J. Comput. Phys.*, 152:457–492, 1999.
- [70] E. Feireisl. Introduction to the theory of partial differential equations – Lecture notes. www-m7.ma.tum.de/lehre/pde2004/scriptum/pde.pdf.
- [71] 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.
- [72] A.C. Galeão and E.G Dutra do Carmo. A consistent approximate upwind petrov-galerkin method for convection dominated problems. *Comp. Meth. Appl. Mech. Engrg.*, 68, 1989.
- [73] T. Gallouët, J.-M. Hérard, and N. Seguin. Some approximate godunov schemes to compute shallow-water equations with topography. *Computer and Fluids*, 32:479–513, 2003.
- [74] P.R. Garabedian. *Basic linear partial differential equations*. Academic Press, 1975.
- [75] P.R. Garabedian. *Partial differential equations*. AMS CHELSEA PUBLISHING, Providence, Rhode Island, 1998.
- [76] P. Garcia-Navarro, M.E. Hubbard, and A. Priestley. Genuinely multidimensional upwinding for the 2D shallow water equations. *J. Comp. Phys.*, 121(1):79–93, 1995.
- [77] S. K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47, 1959.

-
- [78] S.K. Godunov. An interesting class of quasi-linear systems. *Dokl. Akad. Nauk.*, 139, 1961.
- [79] L. Gosse. A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. *Comp. Math. Appl.*, 39:135–159, 2000.
- [80] J.M. Greenberg and A.-Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic systems. *SIAM J. Numer. Anal.*, 33:553–582, 1996.
- [81] R.K.S. Hankin. The Euler equations for multi-phase compressible flow in conservation form - Simulation of shock-bubble interactions. *J. Comput. Phys.*, 172:808–826, 2001.
- [82] A. Harten. High-resolution scheme for hyperbolic conservation laws. *J. Comput. Phys.*, 49:357–393, 1983.
- [83] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comput. Phys.*, 49:151–164, 1983.
- [84] A. Harten, J.M. Hyman, and P.D. Lax. On finite-difference approximations and entropy conditions for shocks. *Comm. on Pure and Appl. Math.*, 29:297–322, 1976.
- [85] G. Hauke. A symmetric formulation for computing transient shallow-water flows. *Comp. Meth. Appl. Mech. Engrg.*, 163:111–122, 1998.
- [86] 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.
- [87] J.C.C. Henriques and L.M.C. Gato. A multidimensional upwind matrix distribution scheme for conservative laws. *Computer and Fluids*, 33:755–769, 2004.
- [88] H. Holden, K.-A. Lie, and N. H. Risebro. An unconditionally stable method for the Euler equations. *J. Comput. Phys.*, 150:76–96, 1999. on line at <http://www.math.ntnu.no/~andreas/fronttrack/papers.html>.
- [89] 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.
- [90] 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.
- [91] M.E. Hubbard. A survey of genuinely multidimensional upwinding techniques. Numerical analysis report 7/93, Department of mathematics, University of Reading, 1993.
- [92] M.E. Hubbard and M.J. Baines. Conservative multidimensional upwinding for the steady two-dimensional shallow-water equations. *J. Comput. Phys.*, 138:419–448, 1997.

- [93] M.E. 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.
- [94] 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.
- [95] 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.
- [96] 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.
- [97] 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.
- [98] T.J.R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for CFD II: beyond SUPG. *Comp. Meth. Appl. Mech. Engrg.*, 54:341–355, 1986.
- [99] E. Issman. *Implicit solution strategies for compressible flow equations on unstructured meshes*. PhD thesis, Université Libre de Bruxelles, 1997.
- [100] E. Issman, G. Degrez, and H. Deconinck. Implicit upwind residual-distribution Euler and Navier-Stokes solver on unstructured meshes. *AIAA Journal*, 34:2021–2028, 1996.
- [101] S. Jin and J.-G. Lin. The effects of numerical viscosities I – slowly moving shocks. *J. Comput. Phys.*, 126:373–389, 1996.
- [102] C. Johnson. *Numerical Solution of Partial Differential equations by the Finite Element method*. Cambridge University Press, Cambridge, 1987.
- [103] C. Johnson. The streamline diffusion finite element method for compressible and incompressible flow. *VKI LS 1990-05, von Karman Institute for Fluid Dynamics*, 1990.
- [104] H.O. Kreiss. Difference approximations for initial-boundary value problems for hyperbolic differential equations. In D. Greenspan, editor, *Numerical solutions of nonlinear partial differential equations*, pages 140–166. Wiley, New York, 1966.
- [105] S.N. Kruzkov. First order quasi-linear equations in several independent variables. *Math. USSR Sbornik*, 10:217–243, 1970.
- [106] A. Kurganov and E. Tadmor. Solution of two-dimensional Riemann problems without Riemann solvers. *Numerical Methods for Partial Differential Equations*, 18:548–608, 2002.
- [107] J.O. Langseth and R.J. LeVeque. Wave propagation method for three-dimensional hyperbolic conservation laws. *J. Comput. Phys.*, 165:126–166, 2000.

-
- [108] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Masson, Paris, France, 1986.
- [109] A. Lerat and C. Corre. A residual-based compact scheme for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 170:642–675, 2001.
- [110] R.J. LeVeque. *Numerical methods for conservation laws*. Birkhäuser, 1992.
- [111] R.J. LeVeque. Wave propagation algorithms for multi-dimensional hyperbolic systems. *J. Comput. Phys.*, 131:327–353, 1997.
- [112] R.J. LeVeque. Balancing source terms and flux gradients in high-resolution godunov method: the quasi-steady wave propagation algorithm. *J. Comp. Phys.*, 146:346–365, 1998.
- [113] 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.
- [114] T.E. Magin. *Modeling and computations of inductive plasma flows*. PhD thesis, Université Libre de Bruxelles, 2004.
- [115] T.E. Magin and G. Degrez. Transport algorithms for partially ionized unmagnetized plasmas. *J. Comput. Phys.*, 198:424–449, 2004.
- [116] T.E. Magin and G. Degrez. Transport properties for partially ionized unmagnetized plasmas. *Physical Review E*, 70, 2004.
- [117] L. Mesaros. *Multi-dimensional Fluctuation-Splitting Schemes for the Euler equations on unstructured grids*. PhD thesis, University of Michigan, 1995.
- [118] 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.
- [119] M. Mezine and R. Abgrall. Upwind multidimensional residual schemes for steady and unsteady flows. In *ICCFD2 International Conference on Computational Fluid Dynamics 2*, pages 165–170, Sidney, Australia, July 2002.
- [120] 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, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [121] M.S. Mock. Systems of conservation laws of mixed type. *J. Diff. Eqns.*, 37:70–88, 1980.
- [122] H. Nishikawa, M. Rad, and P.L. Roe. Grids and solutions from residual minimization. In *ICCFD Proceedings*. Springer-Verlag, 2000.
- [123] 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.

- [124] H. Nishikawa and P.L. Roe. On high-order fluctuation splitting schemes for Navier-Stokes equations. In *ICCFD3 Proceedings*. Springer-Verlag, 2004.
- [125] S. Osher. Convergence of generalized MUSCL schemes. *SIAM J. Numer. Anal.*, 22:947–961, 1985.
- [126] H. Paillère. *Multidimensional Upwind residual Discretization Schemes for the Euler and Navier-Stokes Equations on Unstructured Meshes*. PhD thesis, Université Libre de Bruxelles, 1995.
- [127] H. Paillère, J.-C. Carette, and H. Deconinck. Multidimensional upwind and SUPG methods for the solution of the compressible flow equations on unstructured grids. VKI-LS 1994-05, 1994. Computational Fluid Dynamics.
- [128] 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.
- [129] H. Paillère and H. Deconinck. Compact cell-vertex convection schemes on unstructured meshes. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 1–49. Vieweg, Braunschweig, 1997.
- [130] H. Paillère and H. Deconinck. Multidimensional upwind residual distribution schemes for the 2D Euler equations. In H. Deconinck and B. Koren, editors, *Euler and Navier-Stokes Solvers Using Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 51–112. Vieweg, Braunschweig, 1997.
- [131] 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.
- [132] 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.
- [133] B. Perthame. Convergence of the n scheme for linear advection equations. In J. Rodriguez, editor, *Proc. of SIAM’94 Conference*. Lisbon, 1994.
- [134] 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*, pages 88–93. Springer Verlag, 2002.
- [135] J.J. Quirk. A contribution to the great Riemann solver debate. *International Journal for Numerical Methods in Fluids*, 18:555–574, 1994.
- [136] M. Rad and P.L. Roe. An Euler code that can compute potential flow. In *Proc. 2nd Int. Symposium on FV*. Hermes, 1999.
- [137] M. Renardy and R. Rogers. *An introduction to partial differential equations*. Springer-Verlag, Heidelberg, 2004.

-
- [138] M. Ricchiuto, R. Abgrall, and H. Deconinck. Compact conservative residual discretizations of the shallow-water equations on unstructured grids. In preparation.
- [139] M. Ricchiuto, R. Abgrall, and H. Deconinck. Construction of very high order residual distribution schemes for unsteady advection: preliminary results. *VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [140] M. Ricchiuto, Á. Csík, and H. Deconinck. Space-time residual distribution schemes and application to unsteady two-phase computations on unstructured meshes. Technical Report VKI-PR 2001-23, von Karman Institute for Fluid Dynamics, 2001.
- [141] 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, 2004.
- [142] M. Ricchiuto, Á. Csík, and H. Deconinck. Residual distribution for general time dependent conservation laws. *J. Comput. Phys.*, 209(1):249–289, 2005.
- [143] M. Ricchiuto and H. Deconinck. Time-accurate solution of hyperbolic partial differential equations using FCT and residual distribution. Technical Report VKI-SR 99-33, von Karman Institute for Fluid Dynamics, 1999.
- [144] 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.
- [145] M. Ricchiuto, J. Dobeš, R. Abgrall, and H. Deconinck. Space-time residual distribution schemes for time dependent viscous laminar flows. In preparation.
- [146] M. Ricchiuto, N. Villedieu, R. Abgrall, and H. Deconinck. Very high-order residual distribution schemes for advection-diffusion. In preparation.
- [147] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43:357–372, 1981.
- [148] P. L. Roe. Linear advection schemes on triangular meshes. Technical Report CoA 8720, Cranfield Institute of Technology, 1987.
- [149] P. L. Roe. “Optimum” upwind advection on a triangular mesh. Technical report, ICASE, NASA Langley R.C., 1990.
- [150] 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.
- [151] 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.
- [152] E. Süli. An introduction to the numerical analysis of PDEs – Lecture notes. <http://web.comlab.ox.ac.uk/oucl/work/endre.suli/nspde.ps>.

- [153] M. Seaïd. Non-oscillatory relaxation methods for the shallow-water equations in one and two space dimensions. *International Journal for Numerical Methods in Fluids*, 46:457–484, 2004.
- [154] K. Sermeus and H. Deconinck. Solution of the steady Euler and Navier-Stokes equations using residual distribution schemes. *VKI LS 2003-05, 33rd Computational Fluid dynamics Course, von Karman Institute for Fluid Dynamics*, 2003.
- [155] K. Sermeus and H. Deconinck. An entropy fix for multidimensional upwind residual distribution schemes. *Computer and Fluids*, 34(4):617–640, 2005.
- [156] D. Serre. *Systems of conservation laws I - Hyperbolicity, Entropies, Shock waves*. Cambridge University Press, 1999.
- [157] C.-W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In A. Quarteroni, editor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, volume 1697 of *Lecture Notes in Mathematics*, pages 325–432. Springer-Verlag, Heidelberg, 1998.
- [158] C.-W. Shu. High-order methods for computational physics. 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 439–582. Springer-Verlag, Heidelberg, 1999.
- [159] D. Sidilkover. A new time-space accurate scheme for hyperbolic problems I: quasi-explicit case. Technical report, ICASE, 1998.
- [160] D. Sidilkover and P.L. Roe. Unification of some advection schemes in two dimensions. *Technical Report 95-10, ICASE*, 1995.
- [161] S.P. Spekreijse. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comp.*, 49:135–155, 1987.
- [162] R. Struijs. *A Multi-Dimensional Upwind Discretization Method for the Euler Equations on Unstructured Grids*. PhD thesis, University of Delft, Netherlands, 1994.
- [163] R. Struijs and H. Deconinck. Multidimensional upwind schemes for the Euler equations using fluctuation distribution on a grid consisting of triangles. 8th GAMM conference, 1989.
- [164] 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.
- [165] R. Struijs, H. Deconinck, and P.L. Roe. Fluctuation splitting schemes for the 2D Euler equations. VKI-LS 1991-01, 1991. Computational Fluid Dynamics.
- [166] A. Szepessy. *Convergence of the streamline diffusion finite element method for conservation laws*. PhD thesis, University of Göteborg, 1989.
- [167] E. Tadmor. Skew-selfadjoint form for systems of conservation laws. *J. Math. Anal. Appl.*, 103:428–442, 1984.

-
- [168] E. Tadmor. Entropy functions for symmetric systems of conservation laws. *J. Math. Anal. Appl.*, 122:355–359, 1987.
 - [169] E. Tadmor. Approximate solution of nonlinear conservation laws and related equations. In R. Spigler and S. Venakides, editors, *Recent advances in partial differential equations and applications, Proc. 1996 Venice conference in honor of P.D. Lax and L. Nirenberg on their 70th birthday*, volume 54 of *AMS Proceedings Symp. Appl. Math.* 1997.
 - [170] E. Tadmor. Approximate solution of nonlinear conservation laws. In A. Quarteroni, editor, *Advanced numerical approximation of nonlinear hyperbolic equations*, volume 1697 of *Lecture notes in mathematics*. 1998.
 - [171] E. Tadmor. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numerica*, 12:451–512, 2003.
 - [172] G.T. Tomaich. *A genuinely multidimensional upwinding algorithm for the Navier-Stokes equations on unstructured grids using a compact, highly parallelizable spatial discretization*. PhD thesis, University of Michigan, 1995.
 - [173] L. Tosatto, M. Ricchiuto, and H. Deconinck. Advection-reaction equations in the fluctuation splitting framework. VKI SR04-44, 2004.
 - [174] S. Ulbrich. Partial differential equations - Lecture notes. www-m1.ma.tum.de/m1/personen/sulbrich/pdeLMU/notes/pde_ss03.ps.
 - [175] 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.
 - [176] E. van der Weide. *Compressible Flow Simulation on Unstructured Grids using Multi-dimensional Upwind Schemes*. PhD thesis, Delft University of Technology, Netherlands, 1998.
 - [177] 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.
 - [178] 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 Multi-Dimensional Upwind Schemes and Multigrid Acceleration*, volume 57 of *Notes on Numerical Fluid Mechanics*, pages 113–139. Vieweg, Braunschweig, 1997.
 - [179] 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.
 - [180] M. van Dyke. *An album of fluid motion*. The Parabolic Press, Stanford, California, 1982.

- [181] N. Villedieu, M. Ricchiuto, and H. Deconinck. Study of a 3rd order residual distributive scheme for advection-diffusion equations. Technical Report VKI-PR 04-24, von Karman Institute for Fluid Dynamics, 2004.
- [182] Z.J. Wang. Spectral (finite) volume method for conservation laws on unstructured grids: basic formulation. *J. Comput. Phys.*, 178:210–251, 2002.
- [183] Z.J. Wang. Spectral (finite) volume method for conservation laws on unstructured grids II: extension to two-dimensional scalar advection. *J. Comput. Phys.*, 179:665–697, 2002.
- [184] Z.J. Wang. Spectral (finite) volume method for conservation laws on unstructured grids III: extension to one-dimensional systems. *J. Sci. Computing*, 20:137–157, 2004.
- [185] Z.J. Wang. Spectral (finite) volume method for conservation laws on unstructured grids IV: extension to the two-dimensional Euler equations. *J. Comput. Phys.*, 194:716–741, 2004.
- [186] W.A. Wood and W.L. Kleb. Diffusion characteristics of finite volume and fluctuation splitting schemes. *J. Comput. Phys.*, 153:353–377, 1999.
- [187] P.R. Woodward and P. Colella. The numerical simulation of two-dimensional flows with strong shocks. *J. Comput. Phys.*, 54:115–173, 1984.
- [188] Y. Xing and C.-W. Shu. High-order finite difference WENO schemes with the exact conservation property for the shallow-water equations. To appear on *J. Comput. Phys.* Report SC-2004-10, Division of applied mathematics, Brown University: <http://www.dam.brown.edu/scicomp/publications/Reports/Y2004/BrownSC-2004-10.pdf>.
- [189] 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. Submitted to *J. Sci. Computing*. Report SC-2004-08, Division of applied mathematics, Brown University: <http://www.dam.brown.edu/scicomp/publications/Reports/Y2005/BrownSC-2005-08.pdf>.
- [190] S.T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31:335–362, 1979.
- [191] Z.-C. Zhang, S.T.J. Yu, and S.-C. Chang. A space-time conservation element and solution element method for solving the two- and three-dimensional unsteady Euler equations using quadrilateral and hexahedral meshes. *J. Comput. Phys.*, 175:168–199, 2002.