Chapter 19

Residual Distribution Schemes: Foundations and Analysis

Herman Deconinck[1] and Mario Ricchiuto[2]

[1] *von Karman Institute for Fluid Dynamics, Rhode-Saint-Genése, Belgium*
[2] *INRIA Project ScAlApplix, cours de la Libération 351, 33405, Talence Cedex, France*

1 INTRODUCTION

In the present contribution, we describe residual distribution ($\mathcal{RD}$) methods, seen from a rather fundamental point of view. We try to focus on the basic features that distinguish these methods from the more traditional finite-volume ($\mathcal{FV}$) and finite-element ($\mathcal{FE}$) methods, at the same time showing the many links and similarities. We aim to make clear that, after almost 25 years of research, there is a very rich framework, even though it is still far from being fully developed.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 3: *Fluids*.

ISBN: 0-470-84699-2.

Historically, the residual-based discretizations discussed in this contribution have their origin in two different research lines. The first line was concerned with the study of cell-vertex $\mathcal{FV}$ schemes by Hall, Morton, and collaborators in the early 1980s (Rudgyard, 1993). They realized that improved accuracy could be obtained by discretizing the residual operator as a whole, instead of treating the terms in the partial differential equation (PDE) separately, since a careful design of the residual operator could then lead to cancellation of truncation error terms resulting from the different terms. At the same time, the stabilization needed for convection operators could be designed considering multidimensional aspects like diffusion along the streamline, following similar ideas used in finite-element schemes. Independently, Ni proposed, in a landmark paper in 1981, a Lax–Wendroff (LW) $\mathcal{RD}$ scheme (Ni, 1981). Similar ideas are also at the basis of the residual-based schemes developed by Lerat and Corre (2001).

The second line of research was the work of Roe, aiming to mimic key properties of the physics of the PDE by the discretization. In 1982, Roe (1982) proposed an upwind residual distribution framework for the 1D Euler equations under the name of 'fluctuation splitting ($\mathcal{FS}$)', starting from a reinterpretation of his flux difference splitting $\mathcal{FV}$ scheme. In 1D, the residual (or *fluctuation* as Roe called it) is just the flux balance (or flux difference) over the cell. Roe's classical $\mathcal{FV}$ scheme (Roe, 1981) corresponds to a downwind distribution of the residual, used to update the solution located at the vertices. The $\mathcal{RD}$ view of Roe's 'first-order' upwind scheme allowed to obtain *second*-order accuracy at steady state on nonuniform grids in the presence of a source term, if this was included in the residual (Roe, 1986a).

Generalization for a scalar convection equation in two space dimensions followed in 1987, with the 'fluctuation' being defined as the flux contour integral, that is, the residual, computed over triangular cells (Roe, 1987). In this paper, some of the most used linear multidimensional upwind ($\mathcal{MU}$) distribution schemes (N and low diffusion A (LDA)) were already introduced. The nonlinear positive (local extremum diminishing (LED)) and second-order version of the N scheme (positive streamwise invariance (PSI) scheme) followed in 1990; see, for example, Struijs, Deconinck and Roe (1991).

Extension to hyperbolic systems in two space dimensions however proved to be much more difficult. Multidimensional characteristic-based decompositions for the Euler equations were explored in the late 1980s, aiming to decompose the equations in a set of minimally coupled scalar convection equations (Deconinck, Hirsch and Peuteman, 1986), or using simple wave models (Roe, 1986b). These decomposition techniques aimed to include multidimensional physics, like the propagation of entropy and total enthalpy along the streamline in smooth steady flow, or acoustic Riemann invariants along the Mach lines in 2D steady supersonic flow.

The application of such decomposition models was first attempted in a standard $\mathcal{FV}$ context, with modest success (Powell, van Leer and Roe, 1990; Parpia and Michalec, 1993; Van Ransbeeck and Hirsch, 1996; Hirsch, Lacor and Deconinck, 1987): the main problem is that classical upwind finite-volume schemes base their upwinding on a splitting of the normal fluxes crossing each cell face, which introduces implicitly a locally 1D model in the direction of the normal. Instead, the residual (the flux divergence in the limit of vanishing cell size) is independent of the geometry, and it makes more sense to split this quantity in its multidimensional components. Attempts of $\mathcal{RD}$ schemes based on multidimensional splittings have led to some remarkable successes, for example, in the case of full decoupling into scalar convection equations, which is possible in 2D supersonic steady flow (Paillère, Deconinck and Roe, 1995; Paillère, 1995). This allowed the straightforward use of the scalar convection schemes. However, application to subsonic and 3D flow has still not lead to superior schemes that justify the increased complexity, although progress has been achieved by the so-called hyperbolic-elliptic splittings (Nishikawa, Rad and Roe, 2001).

An alternative for the decomposition techniques to handle the system case was the introduction of *matrix* distribution schemes, which are an algebraic generalization of the scalar convection schemes, introduced in van der Weide and Deconinck (1996), Paillère (1995) and van der Weide and Deconinck (1997). This approach, which is applicable to any hyperbolic system, is recalled in the present contribution. For a decoupled system, it is equivalent to the decomposition methods combined with the scalar schemes.

In this paper, we attempt to give a review of the principles of the method, with a 'modern' point of view, which benefits from the knowledge of the fundamental theoretical properties acquired over the years. The layout of the paper proceeds as follows: After introducing some generalities, a very basic prototype residual discretization for a scalar *steady* convection equation is presented in Section 3, and its properties of accuracy, monotonicity, and energy stability are discussed. Then a large number of schemes are cast in this framework, also including some $\mathcal{FV}$ and Petrov–Galerkin (PG) finite-element schemes (like streamline upwind Petrov–Galerkin (SUPG)). Special attention is paid to nonlinear LED schemes and conservation for nonlinear conservation laws. Then, in Section 5, the prototype discretization of Section 3 is generalized to include unsteady terms, with the aim of solving the time-accurate problem. Again, accuracy (also in time), monotonicity, and stability are reviewed, and a number of time-accurate schemes are presented. Finally, in Section 6, the extension to systems is discussed, building on the matrix extension of the scalar schemes, and numerical results are presented for the Euler equations of gas dynamics, for a hyperbolic model for homogeneous two-phase flow, and for the shallow water equations. The contribution ends with summarizing the main achievements and highlighting the ongoing research directions and open issues.

2 GENERALITIES

2.1 Model problem: hyperbolic conservation laws

This paper presents a class of numerical discretizations for the model problem

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{F}}(\mathbf{u}) = \boldsymbol{\mathcal{S}}(\mathbf{u}, x, y, t) \quad \text{on}$$
$$\Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^d \times \mathbb{R}^+ \tag{1}$$

with $\mathbf{u}$ a vector of m conserved quantities, d the number of space dimensions (2 or 3), $\boldsymbol{\mathcal{F}}$ the $m \times d$ tensor of conservative fluxes, $\boldsymbol{\mathcal{S}}$ a vector of m source terms, and $\Omega_T = \Omega \times [0, t_f]$ the space-time domain over which solutions are sought. System (1) is 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

$$\mathbf{u}(x_1, \ldots, x_d, t = 0) = \mathbf{u}_0(x_1, \ldots, x_d) \tag{2}$$

We focus on the two-dimensional case $d = 2$, $\mathcal{F}(\mathbf{u}) = (\mathbf{F}(\mathbf{u}), \mathbf{G}(\mathbf{u}))$, and $\vec{x} = (x_1, x_2) = (x, y)$; however the theory easily extends to three dimensions. We assume (1) to be *hyperbolic*, that is, $\forall\ \vec{\xi} = (\xi_1, \xi_2) \in \mathbb{R}^2$, the matrix

$$K(\vec{\xi}, \mathbf{u}) = \frac{\partial \mathbf{F}(\mathbf{u})}{\partial \mathbf{u}}\xi_1 + \frac{\partial \mathbf{G}(\mathbf{u})}{\partial \mathbf{u}}\xi_2 \tag{3}$$

admits a complete set of real eigenvalues and linearly independent eigenvectors. A lot of information can be obtained from the analysis of the scalar ($m = 1$) counterpart of (1):

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathcal{F}(u) = \mathcal{S}(u, x, y, t) \quad \text{on} \quad \Omega_T \subset \mathbb{R}^2 \times \mathbb{R}^+ \tag{4}$$

The most simple example of scalar conservation law we will consider is the linear advection problem

$$\frac{\partial u}{\partial t} + \vec{a} \cdot \nabla u = \mathcal{S}(x, y) \quad \text{on} \quad \Omega_T \subset \mathbb{R}^2 \times \mathbb{R}^+ \tag{5}$$

obtained with $\mathcal{F} = \vec{a}u$, and $\vec{a} = \left(a_1, a_2\right) \in \mathbb{R}^2$ constant, or such that $\nabla \cdot \vec{a} = 0$, and with $\mathcal{S} = \mathcal{S}(x, y)$.

2.2 Notation: mesh geometry

We are concerned with the construction of algorithms for the approximation of solutions to (1) on unstructured triangular meshes. In this section we introduce some basic notation used throughout the text. We denote by $\mathcal{T}_h$, a triangulation of the spatial domain Ω. The mesh parameter h is a reference element length. We denote by E the generic triangle, whose area is denoted by $|E|$. Given a node $j \in E$, $\vec{n}_j$ denotes the inward pointing vector normal to the edge of E opposite to j, scaled by the length of the edge (Figure 1a). Since E has a closed boundary, one has

$$\sum_{j \in E} \vec{n}_j = 0 \tag{6}$$

Given a node $i \in \mathcal{T}_h$, $\mathcal{D}_i$ denotes the set of triangles containing i. By abuse of notation, we will say that $j \in \mathcal{D}_i$ if node j belongs to an element $E \in \mathcal{D}_i$. For any vertex i of the grid, we 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 (Figure 1b). The area of S_i is

$$|S_i| = \sum_{E \in \mathcal{D}_i} \frac{|E|}{3} \tag{7}$$

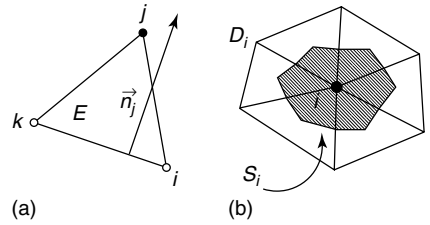

Figure 1. Median dual cell S_i and nodal normal $\vec{n}_j$.

The temporal domain $[0, t_f]$ is discretized by a sequence of discrete time levels $\{t^1 = 0, \ldots, t^n, t^{n+1}, \ldots, t^M = t_f\}$. The schemes we consider allow us to compute an approximation of the solution at time t^{n+1}, known its value at time t^n (and eventually at a finite set of time levels $t^{n-1}, t^{n-2}, \cdots$). Throughout the text, we often focus our attention on a generic space-time slab $\Omega \times [t^n, t^{n+1}]$. The *time-width* of the slab is the time step $\Delta t = t^{n+1} - t^n$.

Throughout the text, the following grid and time-step regularity assumptions are supposed to be true, in any space-time slab $\Omega \times [t^n, t^{n+1}]$:

$$C_1^{\text{mesh}} < \sup_{E \in \mathcal{T}_h} \frac{h}{|E|} < C_2^{\text{mesh}}, \quad C_3^{\text{mesh}} < \frac{\Delta t}{h} < C_4^{\text{mesh}} \tag{8}$$

for some finite, positive constants C_1^{mesh}, C_2^{mesh}, C_3^{mesh}, and C_4^{mesh}.

2.3 Notation: discrete approximation

We consider *continuous* discrete approximations of the unknown, built starting from its values in given locations in the grid. Given a continuous variable $\theta(x, y, t)$, if not stated otherwise, θ_h will denote any continuous discrete approximation of θ, such that given $\theta_i^n = \theta(x_i, y_i, t^n)$, one has $\theta_h(x_i, y_i, t^n) = \theta_i^n$. If θ is a function of the unknown $\mathbf{u}$, we shall often suppose $\theta_h = \theta(\mathbf{u}_h)$. *When specified in the text*, we will use the same notation to refer to continuous piecewise polynomial approximations *in space* of the type

$$\begin{aligned} \theta_h = \theta_h(x, y, t) &= \sum_{i \in \mathcal{T}_h} \psi_i(x, y)\theta_i(t) \\ &= \sum_{i \in \mathcal{T}_h} \psi_i(x, y)\theta(x_i, y_i, t) \\ &= \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \psi_i(x, y)\theta(x_i, y_i, t) \end{aligned} \tag{9}$$

where in general the summation extends not only over the vertices of the triangulation but also over a properly chosen set of nodes placed along the mesh edges and/or within the elements. Several choices are possible to contruct such polynomial approximations. For a review, the reader can refer to Eskilsonn and Sherwin (2005) and references

therein. In any case, we consider basis functions verifying

$$\psi_i(x_j, y_j) = \delta_{ij} \quad \forall i, j \in \mathcal{T}_h, \qquad \sum_{j \in E} \psi_j(x, y) = 1 \quad \forall E \in \mathcal{T}_h \tag{10}$$

with δ_{ij} Kroenecker's delta. In fact, the discrete approximation θ_h is nothing else than a polynomial finite-element interpolant of the values of θ in the chosen set of nodes. In particular, for any element E, we denote by K the number of degrees of freedom (DOF) (nodes) it contains.

Most of the time, however, we will consider continuous piecewise linear approximations of the unknown, and refer to $\{\psi_i\}_{i \in \mathcal{T}_h}$ as the continuous piecewise linear P^1 ($\mathcal{FE}$) basis functions, respecting

$$\psi_i(x_j, y_j) = \delta_{ij} \quad \forall i, j \in \mathcal{T}_h, \qquad \nabla \psi_i|_E = \frac{\vec{n}_i}{2|E|}, \qquad \sum_{j \in E} \psi_j(x, y) = 1 \quad \forall E \in \mathcal{T}_h \tag{11}$$

3 PROTOTYPE DISCRETE APPROXIMATION FOR STEADY PROBLEMS

In this section, we introduce the basics of the $\mathcal{RD}$ approach for steady problems. We recall some elements of the accuracy and stability analysis of the schemes. Examples are given. We focus on the scalar case and, unless stated otherwise, on continuous second-order P1 finite-element approximations built starting from the values of the unknown in the vertices of the grid.

3.1 The residual distribution idea

Consider the solution of the steady limit of (4). We are interested in the following class of discretizations.

Definition 1 (Residual distribution/fluctuation splitting scheme) *Let u_h^0, u_h, $\mathcal{F}_h$, and $\mathcal{S}_h$ be the continuous approximation in space respectively of the initial solution, of the unknown, of the flux, and of the source term. A Residual Distribution or Fluctuation Splitting scheme is defined as a scheme that evolves the nodal values of u_h toward steady state as follows.*

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

$$\phi^E = \int_E (\nabla \cdot \mathcal{F}_h - \mathcal{S}_h)\, \mathrm{d}x\, \mathrm{d}y = \int_{\partial E} \mathcal{F}_h \cdot \hat{n}\, \mathrm{d}l - \int_E \mathcal{S}_h\, \mathrm{d}x\, \mathrm{d}y \tag{12}$$

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

$$\sum_{j \in E} \phi_j^E = \phi^E \tag{13}$$

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

$$\beta_i^E = \frac{\phi_i^E}{\phi^E} \tag{14}$$

one must have by construction

$$\sum_{j \in E} \beta_j^E = 1 \tag{15}$$

3. *$\forall i \in \mathcal{T}_h$ assemble the contributions from all $E \in \mathcal{D}_i$ and evolve u_i in time by integrating the Ordinary Differential Equation (ODE)*

$$|S_i| \frac{\mathrm{d}u_i}{\mathrm{d}t} + \sum_{E \in \mathcal{D}_i} \phi_i^E = 0 \tag{16}$$

Note that the time derivative in (16) has the role of an iterative means to get to the solution of the steady discrete equations:

$$\sum_{E \in \mathcal{D}_i} \phi_i^E = 0 \quad \forall i \in \mathcal{T}_h \tag{17}$$

The properties of the discrete solution are determined by the distribution strategy, that is, by the choice of the split residuals ϕ_j^E, or equivalently by the choice of the β_j^E coefficients. Independently of this choice, however, under reasonable continuity hypothesis on the ϕ_i^E's, and assuming that the consistent approximation of the flux $\mathcal{F}_h$ is continuous, the following LW theorem can be proven (Abgrall, Mer and Nkonga, 2002; Abgrall and Mezine, 2003b).

Theorem 1 (Lax–Wendroff theorem for $\mathcal{RD}$) *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 (12)–(16) verifies*

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

then u is a weak solution of the problem.

We will show that many $\mathcal{FE}$ and $\mathcal{FV}$ schemes can be recast in the $\mathcal{RD}$ formalism. However, before giving examples of particular schemes, we recall general conditions allowing to characterize the accuracy and the stability of the discretization.

In the remaining text, we will omit the superscript ^{E}, when the reference to a generic element E is clear from the context.

3.2 Accuracy of steady $\mathcal{RD}$ discretizations

We consider the issue of the accuracy of the approximation *for steady smooth problems*. Even though second-order schemes are the main focus of this contribution, we give a definition of a kth order accurate scheme, and recall a necessary condition for a $\mathcal{RD}$ scheme to satisfy such definition. We follow Abgrall and Mezine (2003b); Abgrall (2001); and Abgrall and Roe (2003). The analysis is performed for the scalar case, equation (4). The generalization to system (1) is immediate.

The idea is to derive an estimate of how well a scheme reproduces the weak formulation of the problem, in correspondence of a smooth solution. Suppose a solution exists, say w, such that $\nabla \cdot \mathcal{F}(w) = \mathcal{S}(w, x, y)$ in a pointwise manner. Denote by w_h the continuous piecewise polynomial approximation in space of w (cf. equation (9)). We suppose that w_h is of degree $k-1$, and kth order accurate. Consider then the following quantity:

$$TE(w_h) := \sum_{i \in \mathcal{T}_h} \varphi_i \Big(\sum_{E \in \mathcal{D}_i} \phi_i(w_h) \Big) \tag{18}$$

with φ a smooth compactly supported function $\varphi \in C_0^k(\Omega)$, and $\varphi_i = \varphi(x_i, y_i)$. The notation $\phi_i(w_h)$ indicates that the split residuals have been evaluated starting from the continuous interpolant of the nodal values of the exact solution w (and of the exact flux and source term $\mathcal{F}(w)$ and $\mathcal{S}(w, x, y)$). We recall that the superscript ^{E} has been dropped for simplicity. We also recall that, as in (9), the summation in (18) extends not only over the vertices of the grid but also over a properly chosen set of nodes placed along the mesh edges and/or within the elements (cf. Section 2.3).

Note that w_h *is not the numerical solution given by the* $\mathcal{RD}$ *scheme of Definition* 1*, but the* $k-1$ *degree continuous piecewise polynomial approximation of the smooth exact solution* w. Hence, in general $TE(w_h) \neq 0$. The magnitude of this quantity describing how well an interpolant of the exact solution satisfies the discrete equations is what we define as being the truncation error. It gives an estimate on the accuracy of the scheme. In particular we give the following definition.

Definition 2 (kth order accuracy, steady problems). *A* $\mathcal{RD}$ *scheme is said to be* k*th order accurate at steady state if it verifies* $TE(w_h) = \mathcal{O}(h^k)$*, for any smooth exact solution* w*, with* $TE(w_h)$ *given by (18).*

Consider now φ_h, the kth order accurate continuous piecewise polynomial interpolant of φ, constructed starting from the nodal values φ_i (cf. Section 2.3 and (18)). Owing to the regularity of φ and the assumptions (8), we have

$$\begin{aligned} &\|\varphi\|_{L^\infty(\Omega)} < C_1 \\ &|\varphi_i - \varphi_j| \leq \|\nabla\varphi\|_{L^\infty(\Omega)} h < C_2\, h = \mathcal{O}(h) \\ &\|\nabla\varphi_h\|_{L^\infty(\Omega)} < C_3 \end{aligned} \tag{19}$$

for some finite positive constants C_1, C_2, and C_3, eventually depending on the mesh regularity constants in (8).

We now analyze the truncation error (18). We start by rewriting it as

$$TE(w_h) = \sum_{E \in \mathcal{T}_h} \Big(\sum_{i \in E} \varphi_i \phi_i(w_h) \Big) \tag{20}$$

and write the term between brackets as

$$\begin{aligned} \sum_{i \in E} \varphi_i \phi_i(w_h) &= \sum_{i \in E} \varphi_i \phi_i^{\mathrm{G}}(w_h) + \sum_{i \in E} \varphi_i \big(\phi_i(w_h) - \phi_i^{\mathrm{G}}(w_h) \big) \\ &= \int_E \varphi_h \big(\nabla \cdot \mathcal{F}_h(w_h) - \mathcal{S}_h(w_h, x, y) \big)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad + \sum_{i \in E} \varphi_i \big(\phi_i(w_h) - \phi_i^{\mathrm{G}}(w_h) \big) \end{aligned}$$

with $\phi_i^{\mathrm{G}}(w_h)$ the Galerkin residual

$$\begin{aligned} \phi_i^{\mathrm{G}}(w_h) = \phi_i^{\mathrm{G},a}(w_h) - \phi_i^{\mathrm{G},s}(w_h) &= \int_E \psi_i \nabla \cdot \mathcal{F}_h(w_h)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad - \int_E \psi_i \mathcal{S}_h(w_h, x, y)\, \mathrm{d}x\, \mathrm{d}y \end{aligned} \tag{21}$$

defined as the residual weighted with the continuous polynomial basis functions ψ_i given in equation (9). Thus, equation (20) becomes

$$\begin{aligned} TE(w_h) = &\int_\Omega \varphi_h \big(\nabla \cdot \mathcal{F}_h(w_h) - \mathcal{S}_h(w_h, x, y) \big)\, \mathrm{d}x\, \mathrm{d}y \\ &+ \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \varphi_i \big(\phi_i(w_h) - \phi_i^{\mathrm{G}}(w_h) \big) \end{aligned}$$

Since $\sum_{i\in E}\big(\phi_i(w_h) - \phi_i^{\mathrm{G}}(w_h)\big) = \sum_{i\in E}\big(\phi^h(w_h) - \phi^h(w_h)\big) = 0$ (cf. equation (11)), we can write

$$TE(w_h) = \underbrace{\int_\Omega \varphi_h\big(\nabla\cdot\mathcal{F}_h(w_h) - \mathcal{S}_h(w_h,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y}_{I} + \overbrace{\frac{1}{K}\sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E}(\varphi_i-\varphi_j)\big(\phi_i(w_h)-\phi_i^{\mathrm{G}}(w_h)\big)}^{II}$$

where we recall that K is the total number of DOF (nodes) contained in element E. In the last equation, the term I is associated to the error introduced by the choice of the polynomial interpolation in space, while the second term represents the additional error introduced by the $\mathcal{RD}$ discretization.

We first estimate I. Since by hypothesis w verifies (4) in a pointwise manner, we have

$$\begin{aligned}&\int_\Omega \varphi_h\big(\nabla\cdot\mathcal{F}_h(w_h) - \mathcal{S}_h(w_h,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y\\ &\quad= \int_\Omega \varphi_h\big(\nabla\cdot\big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\big)\\ &\qquad- \int_\Omega \varphi_h\big(\mathcal{S}_h(w_h,x,y)-\mathcal{S}(w,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

We decompose the first integral in elemental contributions, use Green–Gauss' formula on each element, and sum up. Owing to the continuity of $\mathcal{F}_h(w_h)$ and φ_h across edges, and the compactness of the support of φ_h, we get

$$\begin{aligned}&\int_\Omega \varphi_h \quad \nabla\cdot\big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\,\mathrm{d}x\,\mathrm{d}y\\ &\quad= -\int_\Omega \nabla\varphi_h\,\big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\,\mathrm{d}x\,\mathrm{d}y = \mathcal{O}(h^k)\end{aligned}$$

provided that $\mathcal{F}_h(w_h)$ is a kth order accurate approximation of $\mathcal{F}(w)$, and thanks to (19). *If $\mathcal{S}_h$ is also a kth order accurate approximation of $\mathcal{S}$*, then (19) ensures that

$$\int_\Omega \varphi_h\big(\mathcal{S}_h(w_h,x,y)-\mathcal{S}(w,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y = \mathcal{O}(h^k) \tag{22}$$

Hence, on a smooth solution, for kth order flux and source term approximations, we have $\mathrm{I} = \mathcal{O}(h^k)$.

Then we estimate II. We start by estimating $\phi_i^{\mathrm{G}}(w_h)$:

$$\begin{aligned}\phi_i^{\mathrm{G}}(w_h) &= \int_E \psi_i\ \nabla\cdot\big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\,\mathrm{d}x\,\mathrm{d}y\\ &\quad- \int_E \psi_i\big(\mathcal{S}_h(w_h,x,y)-\mathcal{S}(w,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y\\ &= \oint_{\partial E}\psi_i\big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\cdot\hat{n}\,\mathrm{d}l\\ &\quad- \int_E \big(\mathcal{F}_h(w_h)-\mathcal{F}(w)\big)\cdot\nabla\psi_i\,\mathrm{d}x\,\mathrm{d}y\\ &\quad+ \int_E \psi_j\big(\mathcal{S}_h(w_h,x,y)-\mathcal{S}(w,x,y)\big)\,\mathrm{d}x\,\mathrm{d}y\\ &= \mathcal{O}(h^{k+1}) + \mathcal{O}(h^{k+1}) + \mathcal{O}(h^{k+2}) = \mathcal{O}(h^{k+1})\end{aligned} \tag{23}$$

having used (19), the fact that $\mathcal{F}_h$ and $\mathcal{S}_h$ are kth order accurate, that $\nabla\psi_i = \mathcal{O}(h^{-1})$ (cf. equation (11)), the boundedness of ψ_i, and the estimates $|\partial E| = \mathcal{O}(h)$, and $|E| = \mathcal{O}(h^2)$. Note that the number of nodes in each element is bounded, while the total number of elements in a regular (in the sense of (8)) triangulation is of $\mathcal{O}(h^{-2})$. Owing to this and to (19), we get for the error

$$\begin{aligned}TE(w_h) &= \mathcal{O}(h^k) + \mathcal{O}(h^{-2})\times\mathcal{O}(h)\times\mathcal{O}(\phi_i(w_h))\\ &\quad+ \mathcal{O}(h^{-2})\times\mathcal{O}(h)\times\mathcal{O}(\phi_i^{\mathrm{G}}(w_h))\\ &= \mathcal{O}(h^k) + \mathcal{O}(h^{-1})\times\mathcal{O}(\phi_i(w_h))\end{aligned}$$

where the quantity $\mathcal{O}(\phi_i(w_h))$ denotes rather the supremum over the nodes and over the elements of the magnitude of the split residuals.

At last we have an error estimate allowing to formulate a necessary condition for second order of accuracy.

Proposition 1. *Given a smooth function $\varphi \in C_0^k(\Omega)$, satisfying the regularity assumptions* (19). *Given a triangulation satisfying the regularity assumption* (8). *Given w_h, the $k-1$ degree, k*th *order accurate continuous piecewise polynomial interpolant of w, a smooth exact solution to* (4)*, and denoting by $\mathcal{F}_h$ and $\mathcal{S}_h$ continuous k*th *order accurate approximations to the exact flux and source term $\mathcal{F}(w)$, and $\mathcal{S}(w,x,y)$ on $\mathcal{T}_h$. Then, in two space dimensions, a $\mathcal{RD}$ scheme verifies the truncation error estimate*

$$TE(w_h) := \sum_{i\in\mathcal{T}_h}\varphi_i\sum_{E\in\mathcal{D}_i}\phi_i(w_h) = \mathcal{O}(h^k) \tag{24}$$

provided that the following condition is met:

$$\phi_i(w_h) = \mathcal{O}(h^{k+1}) \quad \forall\, i\in E \text{ and } \forall\, E\in\mathcal{T}_h \tag{25}$$

Condition (25) guarantees that *formally the scheme has an $\mathcal{O}(h^k)$ error*. In practice, there is no guarantee that this convergence rate is observed, unless some other (stability) constraints are respected. For example, even though it does verify the accuracy condition, the Galerkin scheme (21) is known to be unstable when applied to (4), and to diverge

when the mesh is refined. In this sense the condition of Proposition 1 is only necessary.

Linearity or k-exactness preservation. The last proposition allows us to introduce an important class of schemes. Given continuous and kth order accurate flux and source terms approximations, $\mathcal{F}_h$ and $\mathcal{S}_h$, for a smooth exact solution one has

$$\begin{aligned}\phi^E(w_h) &= \int_E (\nabla \cdot \mathcal{F}_h(w_h) - \mathcal{S}_h(w_h))\,\mathrm{d}x\,\mathrm{d}y \\ &= \oint_{\partial E} (\mathcal{F}_h(w_h) - \mathcal{F}(w)) \cdot \hat{n}\,\mathrm{d}l \\ &\quad - \int_E (\mathcal{S}_h(w_h, x, y) - \mathcal{S}(w, x, y))\,\mathrm{d}x\,\mathrm{d}y \\ &= \mathcal{O}(h^{k+1}) + \mathcal{O}(h^{k+2}) = \mathcal{O}(h^{k+1}) \end{aligned} \tag{26}$$

since $|\partial E| = \mathcal{O}(h)$ and $|E| = \mathcal{O}(h^2)$. As a consequence, we can give the following characterization.

Definition 3 (Linearity or k-exactness preserving scheme) *A $\mathcal{RD}$ scheme is linearity preserving ($\mathcal{LP}$) or more generally k-exactness preserving if its distribution coefficients β_j defined in* (14)*, are uniformly bounded with respect to the solution and the data of the problem:*

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

$\mathcal{LP}$ schemes satisfy by construction the necessary condition for kth order of accuracy of Proposition 1.

The term $\mathcal{LP}$, initially introduced to refer to second-order schemes of this type, is used here to denote in general $\mathcal{RD}$ schemes with uniformly bounded distribution coefficients. In fact a better denomination is *k-exactness preservation*, a term introduced by Barth (2003) in the context of $\mathcal{FV}$ schemes to denote schemes based on a kth degree polynomial reconstruction.

Let us now give a few remarks on the choice of $\mathcal{F}_h$. The result of Proposition 1 is valid provided that $\mathcal{F}_h$ is kth order accurate. A simple way to achieve this is to use for $\mathcal{F}_h$ the same polynomial nodal interpolant as used for w_h. This greatly simplifies the computation of the residual, which can be evaluated directly, by computing, once and for all, edge integrals of the shape functions. This approach is similar to the *quadrature-free* implementation of the discontinuous Galerkin method (Atkins and Shu, 1996). Clearly, the more expensive choice $\mathcal{F}_h(u_h) = \mathcal{F}(u_h)$, with u_h as in (9), is also possible.

On the discrete treatment of source terms. The treatment of the source terms deserves particular attention. Not only is it important for a wide variety of applications, but its analysis will also be useful when discussing the extension of the schemes to time-dependent problems.

Note that the condition $\phi^E = \mathcal{O}(h^{k+1})$ is easily shown to be verified with a $k-1$th order accurate source term representaion (cf. equation (26)). This is a bit misleading, since one might conclude that a lower order approximation of the source term would suffice. In reality, for the analysis to be valid, condition (22) must also be satisfied, hence $\mathcal{S}_h$ really needs to be a kth order approximation. A possible choice for $\mathcal{S}_h$ is the same $k-1$ degree piecewise polynomial as used for w_h, ultimately leading again to a quadrature-free algorithm in which the integral of $\mathcal{S}_h$ only depends on integrals of the basis functions, which can be stored in a preprocessing step. Alternatively, one might choose a representation of the type $\mathcal{S}_h = \mathcal{S}(u_h, x, y)$, with u_h as in (9), and chose a proper quadrature formula to integrate it over an element. The accuracy of such formula has to be consistent with estimate (22). For a second-order scheme, for example, one could use

$$\int_E \mathcal{S}_h(w_h, x, y)\,\mathrm{d}x\,\mathrm{d}y = |E|\mathcal{S}_h(w_\mathrm{G}, x_\mathrm{G}, y_\mathrm{G})$$

where w_G, x_G, y_G denote the solution value and coordinates of the gravity center of element E.

Condition (25) can also be used to show that, in general, pointwise discretizations of the source term are only first-order accurate. *We focus on the case of a piecewise linear approximation of the unknown*, and consider the following variant of (16)

$$|S_i|\frac{\mathrm{d}u_i}{\mathrm{d}t} + \sum_{E\in\mathcal{D}_i} \beta_i \phi^{E,a} = |S_i|\mathcal{S}(u_i, x_i, y_i) \tag{27}$$

where $\phi^{E,a}$ denoted the *advective* element residual

$$\phi^{E,a} = \int_E \nabla \cdot \mathcal{F}_h(u_h)\,\mathrm{d}x\,\mathrm{d}y$$

which, without loss of generality, we have assumed to be distributed by means of a $\mathcal{LP}$ scheme. Equation (27) is the semidiscrete nodal approximation obtained when the forcing term is approximated in a pointwise fashion. Scheme (27) is obtained with the following definition of split residuals:

$$\begin{aligned}\phi_i(u_h) &= \beta_i \phi^{E,a}(u_h) - \frac{|E|}{3}\mathcal{S}_i \\ &= \int_E \left(\beta_i \nabla \cdot \mathcal{F}_h(u_h) - \frac{1}{3}\mathcal{S}_i\right)\mathrm{d}x\,\mathrm{d}y \end{aligned} \tag{28}$$

having set $\mathcal{S}_i = \mathcal{S}(u_i, x_i, y_i)$. Let us now consider a smooth exact solution of the problem w, and let us check what

condition (25) becomes for scheme (28). First of all, we note that

$$\sum_{j\in E}\phi_j(u_h) = \phi^{E,a}(u_h) - \frac{|E|}{3}\sum_{j\in E}\mathcal{S}_j$$
$$= \phi^{E,a}(u_h) - \int_E \mathcal{S}_h \,\mathrm{d}x\,\mathrm{d}y$$

having denoted with $\mathcal{S}_h$ the piecewise linear approximation of the source term obtained with (9). This approximation is second order in smooth areas. Hence, with this approach we can hope, at most, to reach second order of accuracy. In particular, let us also assume that $\mathcal{F}_h$ is second order, and let us now estimate the split residual (28) when it is evaluated on a discrete interpolant of the smooth exact solution w:

$$\begin{aligned}\phi_i(w_h) &= \int_E \left(\beta_i \nabla\cdot\mathcal{F}_h(w_h) - \frac{1}{3}\mathcal{S}_i\right)\mathrm{d}x\,\mathrm{d}y\\ &= \int_E \Big(\beta_i \nabla\cdot\mathcal{F}_h(w_h) - \beta_i\nabla\cdot\mathcal{F}(w)\\ &\quad + \beta_i\mathcal{S}(w,x,y) - \frac{1}{3}\mathcal{S}_i\Big)\mathrm{d}x\,\mathrm{d}y\\ &= \beta_i\phi^{E,a}(w_h - w) + \int_E\left(\beta_i\mathcal{S}(w,x,y) - \frac{1}{3}\mathcal{S}_i\right)\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

having used the fact that $\nabla\cdot\mathcal{F}(w) - \mathcal{S}(w,x,y) = 0$. One easily sees that $\phi^{E,a}(w_h - w) = \mathcal{O}(h^3)$, and hence for (25) to be verified in the case of second-order accuracy, we must have

$$I^s = \int_E \left(\beta_i\mathcal{S}(w,x,y) - \frac{1}{3}\mathcal{S}_i\right)\mathrm{d}x\,\mathrm{d}y = \mathcal{O}(h^3) \qquad (29)$$

One immediately recognizes that, apart from the unstable central scheme obtained with the choice $\beta_i = 1/3$, $\forall\, i \in E$ (cf. Section 4), in general *condition (29) is never respected*. In particular, the boundedness of $\mathcal{S}$ and of $\mathcal{S}_i$ leads in the general case to the estimate $I^s = \mathcal{O}(h^2)$.

Proposition 2 ($\mathcal{LP}$ schemes and pointwise source term) *Apart from the centered scheme obtained with $\beta_i^{\mathrm{C}} = 1/3$, $\forall\, i \in E$, a $\mathcal{LP}$ distribution of the advective fluctuation $\phi^{E,a}$ coupled with a pointwise source treatment leads in general to a first-order accurate discretization. The centered scheme is formally second-order accurate.*

Numerical evidence to support the previous analysis and last proposition can be found in Ricchiuto and Deconinck (2002).

As a final remark, we note that one might think of constructing schemes that still approximate the source term in a pointwise manner by adding a proper correction to a $\mathcal{LP}$ scheme:

$$\phi_i = \beta_i\phi^{E,a} - \frac{|E|}{3}\mathcal{S}_i + \Gamma_i$$

By repeating the above accuracy analysis, using the fact that on E we have $\mathcal{S} = \mathcal{S}_i + \nabla\mathcal{S}\big|_i\cdot(\vec{x}-\vec{x}_i) + \mathcal{O}(h^2)$, and that for the exact solution w one has $\nabla\mathcal{S}\big|_i\cdot(\vec{x}-\vec{x}_i) = \nabla(\nabla\cdot\mathcal{F}(w))\big|_i\cdot(\vec{x}-\vec{x}_i)$, it is easy to show that a second-order scheme can still be obtained provided that $\beta_i = 1/3$, and that Γ_i is at least a second-order approximation of

$$\Gamma_i \approx -\int_E \nabla(\nabla\cdot\mathcal{F}(w))\big|_i\cdot(\vec{x}-\vec{x}_i)\,\mathrm{d}x\,\mathrm{d}y$$

Even though it is possible in theory, this technique has never been investigated in literature. The main issues are how to choose Γ_i in practice, and what the stability properties of the final discretization would be.

3.3 Monotonicity: positive cell-vertex schemes on unstructured grids

We now consider the issue of the nonoscillatory character of the approximation. This characterization is achieved by resorting to the theory of positive coefficient discretizations. *We focus on the homogeneous case $\mathcal{S} = 0$, and only consider continuous piecewise linear approximations, for which the discrete unknowns are the values of the solution in the vertices of the triangulation*. We assume that we will be able to write the split residuals as

$$\phi_i = \sum_{\substack{j\in E\\ j\neq i}} c_{ij}(u_i - u_j) \qquad (30)$$

such that the semidiscrete form of the scheme reads (cf. equations (12)–(16))

$$|S_i|\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij}(u_i - u_j), \quad \forall i \in \mathcal{T}_h \qquad (31)$$

This is certainly possible in the case of the scalar advection problem (5), and it is still admissible in the general nonlinear case if (4) can be locally replaced by a properly linearized version of its quasi-linear form

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

For a linearly varying discrete solution, an example of such an admissible linearization is the *conservative mean-value*

flux Jacobian

$$\widetilde{a} = \frac{1}{|E|}\int_E \vec{a}(u_h)\,\mathrm{d}x\,\mathrm{d}y$$

The analysis reported here has the objective of giving conditions on coefficients c_{ij} in (30), which guarantee the existence of a discrete maximum principle for the discrete solution. The first part of the analysis is an adaptation to the case of the $\mathcal{RD}$ method of the LED principles also used in Barth (2003) and Barth and Ohlberger (2004) (see also Spekreijse, 1987) for the analysis of $\mathcal{FV}$ discretizations on unstructured meshes. We then use these principles to present a maximum principle analysis of schemes (12)–(16), when the time derivative is integrated using a two-step scheme.

3.3.1 LED schemes and discrete maximum principle

We start by recalling the LED principle.

Proposition 3 (LED property) *The prototype scheme* (31) *is LED, that is, in the solution of the ODE* (16) *local maxima are nonincreasing and local minima are nondecreasing, if*

$$\widetilde{c}_{ij} = \sum_{E\in\mathcal{D}_i\cap\mathcal{D}_j} c_{ij} \geq 0, \quad \forall j \in \mathcal{D}_i,\ j\neq i \ \text{ and } \ \forall i \in \mathcal{T}_h \tag{32}$$

Proof. From property (32) it follows that

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

is ≤ 0 *if* u_i *is a local maximum* ($u_i \geq u_j$)*, and it is* ≥ 0 *if* u_i *is a local minimum* ($u_i \leq u_j$)*. Hence the result.* □

The LED property guarantees that local extrema are kept bounded by the numerical scheme. A stronger requirement is obtained by requiring each c_{ij} to be positive, leading to a *subelement LED* property:

Corollary 1 (Subelement LED) *Scheme (31) is LED if* $c_{ij} \geq 0 \quad \forall j \in E \ \text{ and } \ \forall E \in \mathcal{D}_i$.

In order to obtain an estimate on the discrete solution, fully discrete versions of (31) need to be analyzed. Here, we consider the following two-level explicit and implicit time discretizations: explicit (forward) Euler (FE), implicit (backward) Euler (BE), Crank–Nicholson (CN) and trapezium rule. For linear problems, the last two are equivalent. The fully discrete version of (31) obtained with one of these time-integration schemes can be compactly written introducing the θ-scheme:

$$|S_i|(u_i^{n+1}-u_i^n) = -\Delta t\sum_{E\in\mathcal{D}_i}\left((1-\theta)\phi_i^{\mathrm{FE}} + \theta\phi_i^{\mathrm{BE}}\right) \tag{33}$$

with the forward Euler (FE) and backward Euler (BE) contributions given by

$$\begin{aligned}\phi_i^{\mathrm{FE}} &= \sum_{\substack{j\in E\\ j\neq i}}\left[c_{ij}(u_i-u_j)\right]^n\\ \phi_i^{\mathrm{BE}} &= \sum_{\substack{j\in E\\ j\neq i}}\left[c_{ij}(u_i-u_j)\right]^{n+1}\end{aligned} \tag{34}$$

The forward and backward Euler schemes, and the trapezium scheme (equivalent to the $\mathcal{CN}$ scheme for linear advection) schemes are obtained from (33) for $\theta = 0$, $\theta = 1$, and $\theta = 1/2$, respectively. Denoting by U^n and U^{n+1} the arrays containing the nodal values of u at time t^n and t^{n+1}, the θ-scheme can be recast in the form:

$$\mathcal{A}\,U^{n+1} = \mathcal{B}\,U^n \tag{35}$$

where the matrices $\mathcal{A}$ and $\mathcal{B}$ are sparse with a fill-in pattern given by the connectivity graph of the grid. The entries of these matrices depend on the c_{ij} coefficients, the time step, and S_i:

$$\begin{aligned}\mathcal{A}_{ii} &= |S_i| + \theta\Delta t\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij}\\ \mathcal{A}_{ij} &= -\theta\Delta t\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij}\\ \mathcal{B}_{ii} &= |S_i| - (1-\theta)\Delta t\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij}\\ \mathcal{B}_{ij} &= (1-\theta)\Delta t\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij}\end{aligned} \tag{36}$$

We have the following result.

Proposition 4 (Positivity — discrete maximum principle) *The space-time discrete analog of* (4) *in the time interval* $[t^n,\ t^{n+1}]$ *represented by the* θ*-scheme* (33)*, verifies*

the global discrete space-time maximum principle

$$u^n_{\min} = \min_{j\in\mathcal{T}_h} u^n_j \le u^{n+1}_i \le \max_{j\in\mathcal{T}_h} u^n_j = u^n_{\max} \tag{37}$$

and the local discrete space-time maximum principle given by

$$\begin{aligned}\overline{u}_i &= \min\left\{u^n_i,\ \min_{\substack{j\in\mathcal{D}_i\\ j\neq i}}(u^n_j, u^{n+1}_j)\right\} \le u^{n+1}_i \\ &\le \max\left\{u^n_i,\ \max_{\substack{j\in\mathcal{D}_i\\ j\neq i}}(u^n_j, u^{n+1}_j)\right\} = \overline{U}_i\end{aligned} \tag{38}$$

if the LED condition (32) *holds and under the time-step restriction*

$$|S_i| - (1-\theta)\Delta t \sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} c_{ij} \ge 0 \quad \forall i \in \mathcal{T}_h \tag{39}$$

Under the same time-step constraint, the solution obtained with the explicit FE scheme verifies the sharper bounds

$$\widetilde{u}^n_i = \min_{j\in\mathcal{D}_i} u^n_j \le u^{n+1}_i \le \max_{j\in\mathcal{D}_i} u^n_j = \widetilde{U}^n_i \tag{40}$$

In particular, the BE scheme verifies (37) and (38) $\forall \Delta t > 0$*, while the time step allowed by the CN scheme is twice as large as the one allowed by the positivity of the FE scheme.*

Proof. The proof is obtained by noting that the LED condition (32) guarantees that $\mathcal{A}_{ii} \ge 0$ and $\mathcal{A}_{ij} \le 0\ \forall j \neq i$ 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 it is diagonally dominant. This implies that $\mathcal{A}$ is invertible and $\mathcal{A}^{-1}$ is positive (Berman and Plemmons, 1979): $\mathcal{A}^{-1}_{ij} \ge 0\ \forall i, j$. Consider now the array $U_{\min}$ having the same length of U^n and U^{n+1} but with elements all equal to $u^n_{\min}$. Thanks to the time-step restriction (39), we have $\mathcal{B}_{ij} \ge 0\ \forall i, j$. Hence

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

since $u^n_i \ge u^n_{\min}\ \forall i \in \mathcal{T}_h$. Moreover

$$\begin{aligned}\left(\mathcal{B}U_{\min}\right)_i &= \sum_{j\in\mathcal{D}_i} \mathcal{B}_{ij} u^n_{\min} = |S_i| u^n_{\min} \\ &= \sum_{j\in\mathcal{D}_i} \mathcal{A}_{ij} u^n_{\min} = \left(\mathcal{A}U_{\min}\right)_i\end{aligned}$$

Since $\mathcal{A}U^{n+1} = \mathcal{B}U^n$, this shows that $(\mathcal{A}U^{n+1})_i \ge (\mathcal{A}U_{\min})_i$, $\forall i \in \mathcal{T}_h$. The positivity of $\mathcal{A}^{-1} \ge 0$ implies the left inequality in (37). The right inequality is obtained in a similar way.

The local bounds (38) are instead obtained by using the positivity of the $\widetilde{c}_{ij}$ coefficients (32). In fact, given U^{n+1}, one has for a node i

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

Using the positivity of the $\widetilde{c}_{ij}$'s, the time-step restriction (39) and the Definition (38) of $\overline{U}_i$, one has

$$\begin{aligned}&\left(|S_i| + \theta\Delta t \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}\right)u^{n+1}_i \\ &= \left(|S_i| - (1-\theta)\Delta t \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}\right)u^{n+1}_i \\ &\quad + (1-\theta)\Delta t \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}u^n_j + \theta\Delta t \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}u^{n+1}_j \\ &\le \left(|S_i| + \theta\Delta t \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}\right)\overline{U}_i\end{aligned} \tag{41}$$

which gives the right bound in (38). The left bound is obtained in a similar way. For the explicit FE scheme, the sharper bounds (40) are instead readily obtained by noting that

$$\begin{aligned}u^{n+1}_i &= \left(1 - \frac{\Delta t}{|S_i|}\sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}\right)u^n_i + \frac{\Delta t}{|S_i|}\sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}u^n_j \\ &= \overline{c}_{ii}u^n_i + \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \overline{c}_{ij}u^n_j = \sum_{j\in\mathcal{D}_i} \overline{c}_{ij}u^n_j\end{aligned}$$

Bounds (40) are easily verified using the fact that $\overline{c}_{ij} \geq 0$ $\forall\, i\ j$, owing to (32) and (39), and that $\sum_{j\in\mathcal{D}_i} \overline{c}_{ij} = 1$. The last assertion of the proposition is easily checked by comparing the limiting values of the time step obtained by taking $\theta = 0$, $\theta = 1/2$, and $\theta = 1$ in (39). □

Definition 4 (Positive scheme) *A scheme of the form* (33) *respecting Proposition 4 is said to be positive.*

As done for the LED property, we introduce a local form of positivity. First, 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

$$\begin{aligned}
\mathcal{A}^E_{ii} &= \frac{|E|}{3} + \theta\Delta t \sum_{\substack{j\in E\\ j\neq i}} c_{ij}, \quad \mathcal{A}^E_{ij} = -\theta\Delta t \sum_{\substack{j\in E\\ j\neq i}} c_{ij} \\
\mathcal{B}^E_{ii} &= \frac{|E|}{3} - (1-\theta)\Delta t \sum_{\substack{j\in E\\ j\neq i}} c_{ij} \\
\mathcal{B}^E_{ij} &= (1-\theta)\Delta t \sum_{\substack{j\in E\\ j\neq i}} c_{ij}
\end{aligned} \tag{42}$$

With this notation we have the following trivial result.

Proposition 5 (Local positivity — discrete maximum principle) *The space-time discrete analog of* (4) *in the time interval* $[t^n,\ t^{n+1}]$ *represented by the* θ*-scheme* (33) *verifies the global space-time discrete maximum principle* (37) *and the local space-time discrete maximum principle* (38) (*and* (40) *in the explicit case* $\theta = 0$*), if the subelement LED condition holds and under the time-step restriction*

$$\frac{|E|}{3} - (1-\theta)\Delta t \sum_{\substack{j\in E\\ j\neq i}} c_{ij} \geq 0 \quad \forall i \in E \ \text{ and } \ \forall E \in \mathcal{T}_h \tag{43}$$

In particular, the BE scheme verifies (37) *and* (38) $\forall \Delta t > 0$*, while the time-step restriction of the CN scheme is twice less severe than the one guaranteeing the local positivity of the FE scheme.*

Definition 5 (Locally positive scheme) *A scheme verifying Proposition 5 is said to be locally positive.*

The last proposition shows that local positivity implies positivity. 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 monotonicity of the discretization. Unfortunately, it can be shown that, for high-order time-integration schemes, this limitation has a quite general character (Bolley and Crouzeix, 1978). Finally, following Barth (2003) and Barth and Ohlberger (2004), we mention two important corollaries of Proposition 4. The first is that, thanks to the positivity of the $\widetilde{c}_{ij}$ coefficients implied by condition (32), we have the following:

Proposition 6 (Steady-state discrete maximum principle) *Under the hypothesis that the* $\widetilde{c}_{ij}$ *coefficients in* (32) *are all positive, the steady limit of* (33) *verifies the local maximum principle in space given by*

$$\min_{\substack{j\in\mathcal{D}_i\\ j\neq i}} u^*_j \leq u^*_i \leq \max_{\substack{j\in\mathcal{D}_i\\ j\neq i}} u^*_j \tag{44}$$

where the superscript * *denotes the steady limit* $u^*_j = \lim_{n\to\infty} u^n_j$.

The second and more important consequence is that the solution respects at all times the L^∞ stability bounds:

Theorem 2 (L^∞-stability) *If the hypotheses of proposition 4 are verified in all the time slabs* $\{[t^n,\ t^{n+1}]\}$*, with* $n = 0, \ldots, M-1$*, then scheme (33) is* L^∞*-stable and the following bounds hold for its numerical solution:*

$$\min_{i\in\mathcal{T}_h} u^0_i \leq u^n_j \leq \max_{i\in\mathcal{T}_h} u^0_i, \quad \forall i \in \mathcal{T}_h,\ n \in [1, M] \tag{45}$$

3.4 Linear schemes and Godunov's theorem

It is desirable to have a scheme that is both second-order accurate and that respects a discrete maximum principle. It is known that this is not possible, unless the local structure of the solution is somehow monitored by the c_{ij} coefficients. This is formally expressed by the following definition and theorem (Godunov, 1961; Paillère, 1995; Abgrall and Mezine, 2003b).

Definition 6 (Linear scheme) *A scheme of the form* (31) *is said to be linear if all the* c_{ij} *are independent on the numerical solution.*

Theorem 3. *Linear positive RD schemes cannot be more tha first-order accurate.*

3.5 Energy stability

After having characterized the accuracy and monotonicity (L^∞ stability) of the discretization, we consider a different

type of stability, related to the dissipative behavior of the schemes: the energy stability. We focus on the scalar linear case of (5) with $\mathcal{S} = 0$, and on piecewise linear discrete variation of the solution. It is known that the advection equation is characterized by a bound on the L^2 norm of its exact solutions: the energy (Evans, 1998). At the discrete level, this translates into a stability criterion: for stable schemes energy attains its maximum value at $t = 0$, that is, 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 (31). The analysis is inspired by Barth (1996).

We start by rewriting the prototype scheme in the compact vector form

$$D_{|S_i|}\frac{\mathrm{d}U}{\mathrm{d}t} = -\mathcal{C}U \tag{46}$$

and introducing the discrete analog of the energy of the solution

$$\mathcal{E}_h = \frac{U^T\, D_{|S_i|}\, U}{2} = \int_\Omega \mathcal{I}_h \,\mathrm{d}x\,\mathrm{d}y, \quad \mathcal{I} = \frac{1}{2}u^2 \tag{47}$$

with $\mathcal{I}_h$ piecewise linear, as in (9). The stability of the schemes can be characterized by analyzing

$$\frac{\mathrm{d}\mathcal{E}_h}{\mathrm{d}t} = -U^T\frac{\mathcal{C}+\mathcal{C}^T}{2}U = -U^T\, M^{\mathcal{E}_h}\, U \tag{48}$$

We start with giving the following definition.

Definition 7 (Energy stable scheme — semidiscrete form) *The prototype scheme in semidiscrete form* (31) *is energy stable if*

$$\frac{\mathrm{d}\mathcal{E}_h}{\mathrm{d}t} = -U^T\, M^{\mathcal{E}_h}\, U \le 0 \tag{49}$$

It is common experience that schemes yielding monotone numerical solutions, such as LED and positive schemes, also exhibit a *dissipative behavior*, that is, sharp profiles of the solution are smeared as if a viscous diffusion mechanism was present. To characterize our prototype scheme (31) from the energy point of view, we look at the form of the $M^{\mathcal{E}_h}$ matrix. In particular, from (48) and (31) we have

$$\begin{aligned} M^{\mathcal{E}_h}_{ii} &= \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij}, \quad M^{\mathcal{E}_h}_{ij} = -\frac{1}{2}\left(\widetilde{c}_{ij}+\widetilde{c}_{ji}\right)\\ &= -\sum_{E\in\mathcal{D}_i\cap\mathcal{D}_j}\frac{c_{ij}+c_{ji}}{2} \end{aligned}$$

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 positive semidefiniteness, unless the matrix is also irreducibly diagonally dominant (Berman and Plemmons, 1979). In particular, some of the schemes we consider in this paper can be characterized by the following property.

Proposition 7 (Energy stability of LED schemes — semidiscrete case) *A scheme of the form* (31) *verifying the LED condition* (32) *is energy stable in the sense of Definition* 7 *if*

$$\sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ij} = \sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}} \widetilde{c}_{ji} \quad \forall i \in \tau_h \tag{50}$$

Proof. Simple manipulations allow to recast the quadratic form on the right-hand side in (48) as

$$\begin{aligned} U^T\, M^{\mathcal{E}_h}\, U = &\overbrace{\frac{1}{2}\sum_{\substack{i,j\in\mathcal{T}_h\\ \mathcal{D}_i\cap\mathcal{D}_j\neq 0}} (u_i-u_j)\frac{\widetilde{c}_{ij}+\widetilde{c}_{ji}}{2}(u_i-u_j)}^{\geq 0}\\ &+\sum_{i\in\mathcal{T}_h} u_i\Bigg(\sum_{\substack{j\in\mathcal{D}_i\\ j\neq i}}\frac{\widetilde{c}_{ij}-\widetilde{c}_{ji}}{2}\Bigg)u_i \end{aligned}$$

The LED condition (32) guarantees that the first sum is non-negative. Moreover, if condition (50) is verified, the second term in the last equation vanishes, and hence $U^T\, M^{\mathcal{E}_h}\, U \ge 0$, which is the desired result. □

Additional information is obtained by including the temporal discretization in the analysis. For the θ scheme (33) we have the following result.

Proposition 8 (Discrete energy stability — θ scheme) *The family of schemes represented by the* θ *scheme* (33) *verify the following fully discrete energy balance*

$$\begin{aligned} {\mathcal{E}_h}^{n+1} = {\mathcal{E}_h}^{n} - \Delta t\left(\theta U^{n+1}+(1-\theta)U^n\right)^T\\ \times M^{\mathcal{E}_h}\left(\theta U^{n+1}+(1-\theta)U^n\right) - (2\theta-1)\epsilon_h \end{aligned} \tag{51}$$

with the discrete time energy production ϵ_h *given by*

$$\epsilon_h = \tfrac{1}{2}\left(U^{n+1}-U^n\right)^T D_{|S_i|}\left(U^{n+1}-U^n\right) \ge 0$$

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 one preserving the*

dissipation properties of the spatial discretization. For this reason the CN scheme is said to be energy conservative.

Proof. The proof reduces to showing that the balance (51) 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 (33) 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. □

The last proposition shows that 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 stability of the spatial discrete operator. These competitive effects are controlled by the magnitude of the time step. For stable space discretizations, one might then seek a limiting value of Δt for the time discretization guaranteeing the stability of explicit schemes. This study, not undertaken here, can lead sometimes to time-steps constraints for energy stability that are stricter than the ones that have been proved to yield positivity (see e.g. Tadmor, 2003).

4 EXAMPLES OF $\mathcal{RD}$/$\mathcal{FS}$ SCHEMES FOR STEADY ADVECTION

We finally give some examples of $\mathcal{RD}$ schemes. This is done for the case of the advection equation (5). In this case, the element residual ϕ^E can be expressed in a simple analytical form. For u_h and $\mathcal{S}_h$ both piecewise linear, as in (9), and using the properties of the basis functions (11), one easily shows

$$\phi^E = \int_E \left(\vec{a}\cdot\nabla u_h - \mathcal{S}_h\right) \mathrm{d}x\,\mathrm{d}y = \sum_{j\in E} k_j\, u_j - \sum_{j\in E}\frac{|E|}{3}\mathcal{S}_j \tag{52}$$

where k_j denotes the scalar

$$k_j = \tfrac{1}{2}\vec{a}\cdot\vec{n}_j \tag{53}$$

with $\vec{n}_j$ the scaled inward normal of Figure 1. The k_j parameters in (52) can be used as *sensors* to distinguish between downstream and upstream nodes. In particular, $k_j > 0$ only if $\vec{a}$ is oriented as $\vec{n}_j$, and hence only if node j is downstream. Note that, owing to (6), one also has the identity

$$\sum_{j\in E} k_j = 0 \tag{54}$$

With this notation, we will recall in the following section well-known equivalences between cell-vertex first-order $\mathcal{FV}$ schemes, linear $\mathcal{FE}$ schemes, and $\mathcal{RD}$ schemes. The presentation of these more classical methodologies will give additional input for the analysis of discretizations, which can only be constructed in the $\mathcal{RD}$ framework: the $\mathcal{MU}$ schemes. We mainly focus our attention on the homogeneous case $\mathcal{S} = 0$.

4.1 Finite-volume schemes in $\mathcal{FS}$ formalism

On the dual mesh composed of the median dual cells, consider the piecewise constant approximation u_h', with $u_h'|_{S_i} = u_i \ \forall i \in \mathcal{T}_h$. We consider first-order $\mathcal{FV}$ schemes for which the semidiscrete counterpart of (5) in the homogeneous case reads

$$|S_i|\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\oint_{\partial S_i} \mathcal{F}_h(u_h')\cdot\overline{n}\,\mathrm{d}l = -\sum_{l_{ij}\in\partial S_i} H_h(u_i,u_j)\cdot\vec{n}_{ij}$$

where $H(u,v)$ is the $\mathcal{FV}$ numerical flux, respecting $H(u,u) = \mathcal{F}(u)$, l_{ij} is the portion of ∂S_i separating S_i from S_j (see Figure 2a), $\overline{n}_{ij}$ is the exterior unit normal to ∂S_i on l_{ij}, and $\vec{n}_{ij} = |l_{ij}|\overline{n}_{ij}$ is the scaled exterior normal as in Figure 2(b). 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 $\mathcal{D}_i$:

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

The definition of the median dual cell (see Figure 2), and the fact that the hull composed by the edges opposite to i

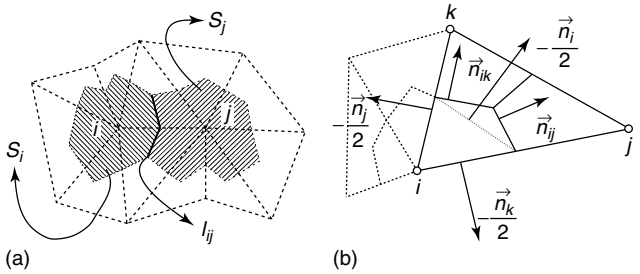

Figure 2. $\mathcal{FV}$ scheme. Neighboring cells S_i and S_j (a) and cell normals (b).

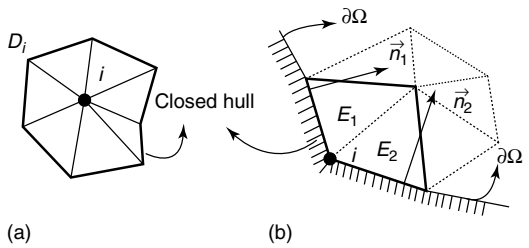

Figure 3. Closed hull around node i.

is closed (see Figure 3), imply the following geometrical identities

$$\sum_{\substack{j\in E\\ j\neq i}} \frac{\vec{n}_j}{2} = -\frac{\vec{n}_i}{2} = \sum_{\substack{j\in E\\ j\neq i}} \vec{n}_{ij} \quad \text{and} \quad \sum_{E\in\mathcal{D}_i} \vec{n}_i = 0 \tag{55}$$

Using these identities, one easily shows that the $\mathcal{FV}$ semidiscrete equation can be equivalently recast as

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

We consider now the family of flux functions defined as

$$H(u_i,u_j) = \frac{\mathcal{F}(u_i)+\mathcal{F}(u_j)}{2}\cdot\vec{n}_{ij} - \frac{1}{2}D(u_i,u_j)(u_j-u_i) \tag{56}$$

with $D(u_i,u_j)$ a *dissipation matrix* (e.g. Roe's absolute value matrix Roe, 1981) 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

$$|S_i|\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\sum_{E\in\mathcal{D}_i}\phi_i, \quad \phi_i = \frac{1}{2}\sum_{\substack{j\in E\\ j\neq i}} ((\mathcal{F}(u_j)-\mathcal{F}(u_i))\cdot\vec{n}_{ij} - D(u_i,u_j)(u_j-u_i)) \tag{57}$$

For the last expression to define a $\mathcal{FS}$ scheme, the ϕ_i's must verify the consistency condition (13), for a *continuous* approximation of the flux. The symmetry of $D(u_i,u_j)$, relation $\vec{n}_{ij} = -\vec{n}_{ji}$, and the first in (55), easily lead to

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

corresponding to (12) integrated exactly for a *continuous piecewise linear approximation of the flux* $\mathcal{F}_h$, as the one obtained with (9), and reducing precisely to (52) for constant (homogeneous) advection. The analogy extends to nonlinear problems, and systems as well (Abgrall, Mer and Nkonga, 2002).

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

$$H(u_i,u_j) = k_{ij}\frac{(u_i+u_j)}{2} - \frac{|k_{ij}|}{2}(u_j-u_i), \quad k_{ij} = \vec{a}\cdot\vec{n}_{ij} \tag{58}$$

using which we finally arrive at the upwind $\mathcal{FV}-\mathcal{RD}$ scheme defined by Paillère (1995)

$$S_i|\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\sum_{E\in\mathcal{D}_i}\phi_i^{\mathcal{FV}-\mathcal{RD}}$$
$$\phi_i^{\mathcal{FV}-\mathcal{RD}} = -\sum_{\substack{j\in E\\ j\neq i}} k_{ij}^-(u_i-u_j) \tag{59}$$

Scheme (59) is of the form (31) with $c_{ij} = -k_{ij}^- \geq 0$, and hence it verifies the subelement LED condition. It verifies Propositions 4, 5, and 6, and Theorem 2, and the related stability bounds. In particular, the time-step restrictions for its positivity and local positivity are given by

$$\Delta t \leq \frac{|S_i|}{(1-\theta)\sum_{E\in\mathcal{D}_i}\sum_{\substack{j\in E\\ j\neq i}} -k_{ij}^-}, \ \forall i\in\mathcal{T}_h$$

and

$$\Delta t \leq \frac{|E|}{3(1-\theta)\sum_{\substack{j\in E\\ j\neq i}} -k_{ij}^-}, \quad \forall E\in\mathcal{D}_i, \ \forall i\in\mathcal{T}_h \tag{60}$$

with $\theta\in[0,1)$. Unconditional positivity is obtained with BE time integration. For this scheme, the distribution coefficients are not explicitly defined. They have to be computed as $\beta_i^{\mathcal{FV}-\mathcal{RD}} = \phi_i^{\mathcal{FV}-\mathcal{RD}}/\phi^E$, and their boundedness for $\phi^E\to 0$ is not guaranteed. Hence, the scheme is not $\mathcal{LP}$. First order of accuracy is observed in practice. However, since $k_{ij} = -k_{ji}$, making use of the first identity in (55), and of the definitions of k_{ij} and k_i, we have

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

Owing to the second relation in (55), for constant advection $\sum_{E\in\mathcal{D}_i} k_i = 0$, which proves that the upwind $\mathcal{FV}$-$\mathcal{RD}$ scheme respects the energy stability criteria of proposition 7. Note that, with reference to Figure 3(b), 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 to the boundary $\vec{n}_1$ and $\vec{n}_2$ scaled by the length of the edges. When included in the energy balance, these terms give an approximation of the energy flux across $\partial\Omega$, the energy balance becoming (see (47) and (48))

$$\frac{\mathrm{d}\mathcal{E}_h}{\mathrm{d}t} = -U^T M^{\mathcal{E}_h} U - \frac{1}{2}\oint_{\partial\Omega}\mathcal{I}_h(\vec{a}\cdot\hat{n})\,\mathrm{d}l$$

with $\hat{n}$ the exterior unit normal to $\partial\Omega$, and $\mathcal{I}_h$. How to handle this extra term is shown in the next section.

4.2 Central and finite-element ($\mathcal{FE}$) schemes

We now consider another family of discretizations, which originally were not formulated as $\mathcal{FS}$ schemes, but that naturally admit a $\mathcal{RD}$ formulation. They are all variations of a central scheme obtained by equidistributing the residual to the nodes of an element. We start by showing the equivalence of this centered $\mathcal{RD}$ scheme with the Galerkin $\mathcal{FE}$ discretization of (5). For steady constant advection, and neglecting the BC terms, the P^1 Galerkin $\mathcal{FE}$ scheme reads

$$\int_\Omega \psi_i\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y = 0 \quad \forall i\in\mathcal{T}_h \tag{61}$$

with ψ_i the linear basis functions (11), and u_h as in (9). In the case of a constant advection speed, using the compactness of the support of the basis functions and (11), the Galerkin scheme can be immediately recast as

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

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

$$\beta_i^{\mathrm{C}} = \frac{1}{3} \tag{62}$$

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

4.2.1 Petrov–Galerkin schemes and streamline dissipation

The Galerkin method is known to be unstable when approximating the advection equation. Consider then the stabilized PG schemes, obtained by adding to the Galerkin discretization a so-called *streamline dissipation* term (Hughes and Brook, 1982; Hughes and Mallet, 1986; Johnson, 1987; Szepessy, 1989):

$$\int_\Omega \psi_i\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y + \underbrace{\sum_{E\in\mathcal{T}_h}\int_E \tau(\vec{a}\cdot\nabla\psi_i)(\vec{a}\cdot\nabla u_h)\,\mathrm{d}x\,\mathrm{d}y}_{\text{PG streamline dissipation}} = 0 \quad \forall i\in\mathcal{T}_h \tag{63}$$

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

$$\begin{aligned} 0 &= \sum_{E\in\mathcal{D}_i}\frac{1}{3}\phi^E + \sum_{E\in\mathcal{D}_i}\tau\frac{k_i}{2|E|}\phi^E \\ &= \sum_{E\in\mathcal{D}_i}\phi_i^{\mathrm{C}} + \sum_{E\in\mathcal{D}_i}\tau\frac{k_i}{2|E|}\phi^E \quad \forall i\in\mathcal{T}_h \end{aligned} \tag{64}$$

which shows the equivalence of stabilized streamline dissipation Galerkin $\mathcal{FE}$ scheme with the class of $\mathcal{LP}$-$\mathcal{RD}$ schemes defined by the distribution coefficient

$$\beta_i^{\text{SD-G}} = \frac{1}{3} + \tau\frac{k_i}{2|E|} \quad \text{with} \quad \tau\geq 0,\ \tau = \mathcal{O}\left(\frac{h}{\|\vec{a}\|}\right) \tag{65}$$

This analogy is of course known for a long time (Carette *et al.*, 1995; Paillère, 1995). However, strictly speaking the analogy is an equivalence only in the constant coefficients case, while in general the $\mathcal{RD}$ and the $\mathcal{FE}$ schemes give different discrete equations since the integrals in (63) no longer reduce to (64). The streamline dissipation terms introduce some kind of upwind bias in the distribution, since we have (see also Section 4.3)

$$\begin{aligned} \beta_i^{\text{SD-G}} &> \beta_i^{\mathrm{C}} \quad \text{if } i \text{ is downstream, hence} \quad k_i > 0 \\ \beta_i^{\text{SD-G}} &< \beta_i^{\mathrm{C}} \quad \text{if } i \text{ is upstream, hence} \quad k_i < 0 \end{aligned}$$

The stabilization mechanism introduced by this upwind bias is better understood by looking at the energy stability of the schemes.

4.2.2 PG schemes: energy stability

Streamline-diffusion $\mathcal{FE}$ schemes have well-known energy stability properties that we will recall here. The energy

balance of the streamline diffusion galerkin (SD-G) scheme reads

$$\frac{d\mathcal{E}_h^{\mathrm{SD-G}}}{\mathrm{d}t} = -\sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E}\frac{1}{3}u_i k_j u_j - \sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E}u_i\frac{k_i\tau k_j}{2|E|}u_j = \frac{d\mathcal{E}_h^{\mathrm{C}}}{\mathrm{d}t} - \epsilon_h^{\mathrm{SD}} \quad (66)$$

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

$$\sum_{i\in E}\sum_{j\in E}\frac{1}{3}u_i k_j u_j = \int_E u_h\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y$$

$$\epsilon_h^{\mathrm{SD}} = \sum_{E\in\mathcal{T}_h}\frac{1}{2|E|}\begin{bmatrix}k_1u_1\\k_2u_2\\k_3u_3\end{bmatrix}^T\begin{bmatrix}\tau&0&0\\0&\tau&0\\0&0&\tau\end{bmatrix}\begin{bmatrix}k_1u_1\\k_2u_2\\k_3u_3\end{bmatrix}\geq 0 \quad (67)$$

showing that the upwind bias of the streamline-diffusion adds an L^2 stabilizing dissipation mechanism. Finally, the global balance can be recast as

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{SD-G}}}{\mathrm{d}t} = -\sum_{E\in\mathcal{T}_h}\int_E u_h\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y - \epsilon_h^{\mathrm{SD}} = -\int_\Omega u_h\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y - \epsilon_h^{\mathrm{SD}} \quad (68)$$

Unless the boundary conditions are taken into account, this only shows that a dissipative mechanism is present, through the ϵ_h^{SD} term. For simplicity, suppose that homogeneous BCs are prescribed. To be completely faithful to the $\mathcal{FE}$ formulation, the BCs should be included in the variational formulation (63) using the admissibility condition (Barth, 1998)

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

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

$$\oint_{\partial\Omega} u_h(\vec{a}\cdot\hat{n})^- u_h\,\mathrm{d}l = 0 \quad (69)$$

either because we impose $u_h = 0$ (inflow boundary, $(\vec{a}\cdot\hat{n})^-\leq 0$), or because $(\vec{a}\cdot\hat{n})^- = 0$ (outflow boundary). We then rewrite the energy estimate (68) as

$$\begin{aligned}\frac{d\mathcal{E}_h^{\mathrm{SD-G}}}{\mathrm{d}t} &= -\int_\Omega u_h\vec{a}\cdot\nabla u_h\,\mathrm{d}x\,\mathrm{d}y - \epsilon_h^{\mathrm{SD}}\\ &= -\frac{1}{2}\oint_{\partial\Omega}u_h(\vec{a}\cdot\hat{n})u_h\,\mathrm{d}l - \epsilon_h^{\mathrm{SD}}\\ &= -\oint_{\partial\Omega}\mathcal{I}(u_h)(\vec{a}\cdot\hat{n})\,\mathrm{d}l - \epsilon_h^{\mathrm{SD}}\\ &= -\oint_{\partial\Omega}\mathcal{I}(u_h)|\vec{a}\cdot\hat{n}|\,\mathrm{d}l - 2\oint_{\partial\Omega}\mathcal{I}(u_h)(\vec{a}\cdot\hat{n})^-\,\mathrm{d}l - \epsilon_h^{\mathrm{SD}}\end{aligned}$$

having used the identity $\vec{a}\cdot\hat{n} = 2(\vec{a}\cdot\hat{n})^- + |\vec{a}\cdot\hat{n}|$, and (47). Using the BCs (69), we obtain the stability estimate

$$\frac{d\mathcal{E}_h^{\mathrm{SD-G}}}{\mathrm{d}t} = -\oint_{\partial\Omega}\mathcal{I}(u_h)|\vec{a}\cdot\hat{n}|\,\mathrm{d}l - \epsilon_h^{\mathrm{SD}}\leq 0 \quad (70)$$

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 (70) (Barth, 1998). The analysis shows that the total energy production can be split into the energy dissipation introduced by the upwind bias (ϵ_h^{SD}) plus the energy production due to the centered discretization. The latter is then simplified taking into account the boundary conditions, finally obtaining an energy stability estimate.

4.2.3 The Rusanov scheme

Among the central schemes, we also report the LED Rusanov's (Rv) scheme (Rusanov, 1961; Abgrall and Mezine, 2003b; Abgrall and Mezine, 2004) defined by

$$\phi_i^{\mathrm{Rv}} = \frac{1}{3}\phi^E + \frac{1}{3}\alpha\sum_{\substack{j\in E\\ j\neq i}}(u_i - u_j),\quad \alpha\geq\max_{j\in E}|k_j| > 0 \quad (71)$$

The Rv scheme is obtained by adding to the centered scheme a stabilizing term. To see this, we rewrite (71) as

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

where (54) has been used in the second equality. The Rv scheme can be recast as in (31) with $3c_{ij} = (\alpha - k_j)\geq 0$ (by definition of α). Hence, the scheme respects the subelement LED condition, and it verifies Propositions 4–6, and Theorem 2. The time-step restrictions for its positivity and

local positivity read

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

and

$$\Delta t \leq \frac{|E|}{(1-\theta)\sum_{\substack{j\in E\\ j\neq i}}(\alpha - k_j)}, \ \forall E \in \mathcal{D}_i, \ \forall i \in \mathcal{T}_h \tag{72}$$

with $\theta \in [0, 1)$. The Rv scheme is unconditionally positive when BE time integration is used in (16). The distribution coefficients of the Rv scheme are not guaranteed to be bounded, and hence the scheme is not $\mathcal{LP}$. The energy stability of the Rv scheme can be easily shown noting that

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{Rv}}}{\mathrm{d}t} = -\sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E}\frac{1}{3}u_i k_j u_j$$
$$-\frac{1}{3}\sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E}u_i\alpha(u_i - u_j) = \frac{\mathrm{d}\mathcal{E}_h^{\mathrm{C}}}{\mathrm{d}t} - \epsilon_h^{\mathrm{Rv}} \tag{73}$$

with the dissipation term reading

$$\epsilon_h^{\mathrm{Rv}} = \sum_{E\in\mathcal{T}_h}\frac{1}{3}\begin{bmatrix}u_1-u_2\\ u_1-u_3\\ u_2-u_3\end{bmatrix}^T\begin{bmatrix}\alpha & 0 & 0\\ 0 & \alpha & 0\\ 0 & 0 & \alpha\end{bmatrix}\begin{bmatrix}u_1-u_2\\ u_1-u_3\\ u_2-u_3\end{bmatrix} \geq 0$$
$$\text{since} \quad \alpha \geq 0, \quad \forall\, E \in \mathcal{T}_h \tag{74}$$

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

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{Rv}}}{\mathrm{d}t} = -\oint_{\partial\Omega}\mathcal{I}(u_h)|a\cdot\hat{n}|\,\mathrm{d}l - \epsilon_h^{\mathrm{Rv}} \leq 0 \tag{75}$$

4.3 A truly multidimensional upwinding strategy

For a linear (or linearized) problem in quasi-linear form, the residual can be expressed as in (52). We consider now the homogeneous case $\mathcal{S} = 0$, and recast (52) in an alternate form. Using the *upwind parameters*

$$k_j^+ = \max(0, k_j), \quad k_j^- = \min(0, k_j) \tag{76}$$

and the identity $k_j = k_j^+ + k_j^-$, one has

$$\phi^E = \sum_{j\in E}k_j^+u_j + \sum_{j\in E}k_j^-u_j$$
$$= \Big(\sum_{j\in E}k_j^+\Big)\Big(\sum_{j\in E}N\,k_j^+u_j + \sum_{j\in E}N\,k_j^-u_j\Big)$$

having introduced the quantity

$$N = \Big(\sum_{j\in E}k_j^+\Big)^{-1} = -\Big(\sum_{j\in E}k_j^-\Big)^{-1} = \frac{1}{2}\Big(\sum_{j\in E}|k_j|\Big)^{-1} > 0 \tag{77}$$

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

$$u_{\mathrm{in}} = \frac{\sum_{j\in E}k_j^-u_j}{\sum_{j\in E}k_j^-} = -\sum_{j\in E}N\,k_j^-u_j$$
$$\text{and} \quad u_{\mathrm{out}} = \sum_{j\in E}N\,k_j^+u_j \tag{78}$$

the residual can be written as (Paillère, 1995)

$$\phi^E = M\,(u_{\mathrm{out}} - u_{\mathrm{in}}), \quad M = \sum_{j\in E}k_j^+ = N^{-1} \tag{79}$$

To give a geometrical interpretation of (79) we observe 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 ζ cutting the triangle (see Figure 4), that is $u_{\mathrm{out}} = u_h(\vec{x}_{\mathrm{out}})$ and $u_{\mathrm{in}} = u_h(\vec{x}_{\mathrm{in}})$, with (Paillère, 1995):

$$\vec{x}_{\mathrm{out}} = \sum_{j\in E}Nk_j^+\vec{x}_j\ , \quad \vec{x}_{\mathrm{in}} = -\sum_{j\in E}Nk_j^-\vec{x}_j$$

Hence, the residual (79) represents a one-dimensional balance along ζ. This framework gives the basis for a truly multidimensional generalization of concepts derived from the study of one-dimensional advection. Depending on how $\vec{a}$ is oriented in E, we can distinguish two situations (see Figure 4). If $\vec{a}$ points in the direction of a single point of E, as in (a), then this point coincides with the outflow point and is the only downstream point. In this situation the element is said to be *1-target*. Conversely, if $\vec{a}$ points in the direction of one of the edges of E, as in (b), then there is only one upstream point coinciding with the inflow point. In this situation, the element is said to be *2-target*. If E is 1-target, then there is a node j such that

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

Similarly, if E is 2-target, then there is a node k such that

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

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 particular, we define the following class of schemes.

Definition 8. *A $\mathcal{FS}$ scheme is multidimensional upwind ($\mathcal{MU}$) if*

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

Clearly, $\mathcal{MU}$ schemes reduce to 1D upwind schemes along the streamline cutting the triangle: all the information contained in the fluctuation is sent to the outflow point. This is an important simplification: *all $\mathcal{MU}$ schemes are equivalent in the 1-target case*, *different $\mathcal{MU}$ schemes are defined just by choosing different distribution strategies in the 2-target case*. The geometrical framework of Figure 4 allows to perform this choice on the basis of heuristics making use of the directional propagation of the information that characterizes exact solutions. There are quite a number of possible choices (Paillère, 1995; Roe, 1987). Here, only two of these will be analyzed in more detail. Before that, we recall the following simple result.

Proposition 9 ($\mathcal{MU}$ schemes, LED, $\mathcal{LP}$ property and energy stability: 1-target case) *In the 1-target configuration, $\mathcal{MU}$ schemes are $\mathcal{LP}$ and positive. Moreover, they are locally dissipative.*

Proof. Let us locally number as $(1, 2, 3)$ the nodes of the 1-target triangle T. Suppose 1 is the only downstream node: $k_1 > 0$, $k_2, k_3 < 0$. One immediately shows that a $\mathcal{MU}$ scheme is $\mathcal{LP}$, by noting that

$$\phi_1 = \phi^E(u_h),\ \phi_2 = \phi_3 = 0 \Longrightarrow \beta_1 = 1,\ \beta_2 = \beta_3 = 0$$

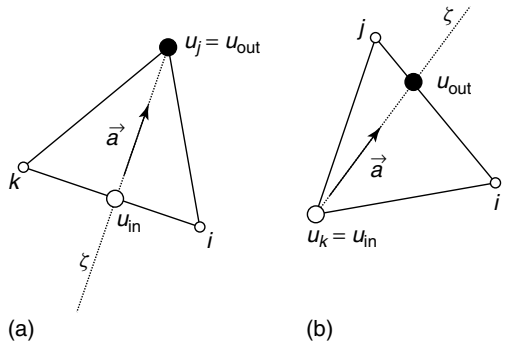

Figure 4. Inflow and outflow state. One-target (a) and two-target element (b).

which are clearly uniformly bounded. Positivity is quickly checked by noting that

$$\phi_1 = \phi^E = \sum_{j\in T} k_j u_j = -k_2(u_1 - u_2) - k_3(u_1 - u_3) = c_{12}(u_1 - u_2) + c_{13}(u_1 - u_3)$$

with $c_{12}, c_{13} \geq 0$ by hypothesis, and having used (54). Lastly, we show that the scheme is locally dissipative. With the notation of (47), we note that we can write for the *total energy of the solution*

$$\begin{aligned}\frac{\partial \mathcal{E}_h}{\partial t} &= \int_\Omega \frac{\partial \mathcal{I}_h}{\partial t} = \int_\Omega \sum_{i\in\mathcal{T}_h} \psi_i \frac{\partial \mathcal{I}_i}{\partial t} = \sum_{i\in\mathcal{T}_h} |S|_i \frac{\partial \mathcal{I}_i}{\partial t} \\ &= \sum_{i\in\mathcal{T}_h} u_i |S|_i \frac{\partial u_i}{\partial t} = -\sum_{i\in\mathcal{T}_h}\sum_{E\in\mathcal{D}_i} u_i \phi_i \\ &= -\sum_{E\in\mathcal{T}_h}\sum_{j\in E} u_j \phi_j = -\sum_{E\in\mathcal{T}_h} \boldsymbol{\Phi}_{\mathcal{I}}^E = -\boldsymbol{\Phi}_{\mathcal{I}} \end{aligned} \tag{80}$$

As seen earlier, a dissipative (viz. energy stable) scheme verifies $\boldsymbol{\Phi}_{\mathcal{I}} = \oint_{\partial\Omega} \mathcal{I}_h \vec{a}\cdot\hat{n}\,\mathrm{d}l + \epsilon_{\mathcal{I}}$ with $\epsilon_{\mathcal{I}} \geq 0$. Hereafter, we show that for a $\mathcal{MU}$ scheme one has in the 1-target case

$$\boldsymbol{\Phi}_{\mathcal{I}}^E = \oint_{\partial E} \mathcal{I}_h \vec{a}\cdot\hat{n}\,\mathrm{d}l + \epsilon_{\mathcal{I}}^{\text{1-target}}, \quad \epsilon_{\mathcal{I}}^{\text{1-target}} \geq 0$$

Recalling that we assumed node 1 to be the only downstream node, a direct calculation shows that

$$\begin{aligned}\epsilon_{\mathcal{I}}^{\text{1-target}} &= \boldsymbol{\Phi}_{\mathcal{I}}^E - \oint_{\partial E} \mathcal{I}_h \vec{a}\cdot\hat{n}\,\mathrm{d}l \\ &= u_1 \phi^E(u_h) - \int_E \vec{a}\cdot\nabla\mathcal{I}_h\,\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{j\in E} u_1 k_j u_j - \sum_{j\in E} \frac{1}{2} u_j k_j u_j = U^T \mathrm{M}_{\text{1-target}} U \\ &= \frac{1}{2} U^T \left(\mathrm{M}_{\text{1-target}} + \mathrm{M}_{\text{1-target}}^T\right) U = U^T \mathrm{M}_{\text{1-target}}^{\text{symm}} U \end{aligned}$$

where U denotes the array $U = [u_1\ u_2\ u_3]$, and with $\mathrm{M}_{\text{1-target}}$ and $\mathrm{M}_{\text{1-target}}^{\text{symm}}$ given by

$$\mathrm{M}_{\text{1-target}} = \frac{1}{2}\begin{bmatrix} k_1 & 2k_2 & 2k_3 \\ 0 & -k_2 & 0 \\ 0 & 0 & -k_3 \end{bmatrix}$$

$$\mathrm{M}_{\text{1-target}}^{\text{symm}} = \frac{1}{2}\begin{bmatrix} k_1 & k_2 & k_3 \\ k_2 & -k_2 & 0 \\ k_3 & 0 & -k_3 \end{bmatrix}$$

We see that $\mathrm{M}^{\mathrm{symm}}_{1\text{-target}}$ has positive diagonal and nonpositive off-diagonal entries; moreover, the row and column sums of its elements are zero owing to (54). Hence, $\mathrm{M}^{\mathrm{symm}}_{1\text{-target}}$ is positive semidefinite (Berman and Plemmons, 1979). As a consequence $\epsilon_{\mathcal{I}}^{1\text{-target}} = U^T \mathrm{M}^{\mathrm{symm}}_{1\text{-target}} U \geq 0$, which is the desired result. □

In the 1-target case, $\mathcal{MU}$ schemes have all the properties one can possibly desire. Note that the last proposition is not in contradiction with Godunov's theorem, since the LED property would require the positivity of the coefficients in all the elements of the mesh, which obviously (and unfortunately) 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 can be retained in the 2-target case by a linear $\mathcal{MU}$ scheme. Two examples are recalled in the following.

4.3.1 *Multidimensional upwind schemes: the LDA scheme*

The LDA is the linear $\mathcal{LP}$-$\mathcal{MU}$ scheme defined by the distribution coefficients:

$$\beta_i^{\mathrm{LDA}} = k_i^+ N = k_i^+ \left(\sum_{j \in E} k_j^+ \right)^{-1} \in [0,\ 1] \qquad (81)$$

In the homogeneous case, (79) gives for the local nodal residuals

$$\phi_i^{\mathrm{LDA}} = \beta_i^{\mathrm{LDA}} \phi^E = k_i^+ (u_{\mathrm{out}} - u_{\mathrm{in}}) \qquad (82)$$

The scheme is clearly $\mathcal{LP}$, since β_i^{LDA} is bounded independently on ϕ^E, and hence $\phi_i = \mathcal{O}(h^3)$. One easily checks that it is not LED (Paillère, 1995). In the 2-target case, a simple geometrical interpretation is possible (Paillère, 1995; Roe, 1987). With reference to Figure 5, we define the subtriangles T_{423} and T_{143}. Simple trigonometry shows that

$$|T_{423}| = \frac{l_{34} k_1}{\|\vec{a}\|},\ |T_{143}| = \frac{l_{34} k_2}{\|\vec{a}\|}$$

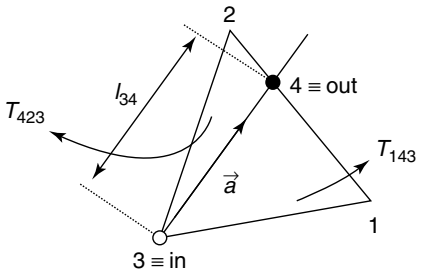

Figure 5. Geometry of $\mathcal{FS}$ schemes. LDA in the two-target case.

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

The distribution coefficients of the two downstream nodes 1 and 2 can be written as the area ratios

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

LDA scheme: energy stability

Using the definition of ϕ_i^{LDA}, and equations (52) and (79), the energy balance of the LDA scheme is

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{LDA}}}{\mathrm{d}t} = -\sum_{E \in \mathcal{T}_h} \sum_{j \in E} u_j \phi_j^{\mathrm{LDA}} = -\sum_{E \in \mathcal{T}_h} u_{\mathrm{out}} M (u_{\mathrm{out}} - u_{\mathrm{in}})$$

Simple manipulations lead to the more convenient expression

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{LDA}}}{\mathrm{d}t} = -\sum_{E \in \mathcal{T}_h} \frac{(u_{\mathrm{out}} + u_{\mathrm{in}})}{2} M (u_{\mathrm{out}} - u_{\mathrm{in}}) - \epsilon_h^{\mathrm{LDA}} \qquad (83)$$

with

$$\epsilon_h^{\mathrm{LDA}} = \frac{1}{2} \sum_{E \in \mathcal{T}_h} (u_{\mathrm{out}} - u_{\mathrm{in}}) M (u_{\mathrm{out}} - u_{\mathrm{in}}) \geq 0 \qquad (84)$$

As for the PG scheme, the energy production of the LDA can be split into a stabilizing term, owing to the upwinding, plus a centered term. In this case, both contributions act along the streamline, making the analysis less clear. Indeed, we can express the energy balance of the LDA scheme as (see equation (47))

$$\frac{d\mathcal{E}_h^{\mathrm{LDA}}}{\mathrm{d}t} = -\sum_{E \in \mathcal{T}_h} M (\mathcal{I}(u_{\mathrm{out}}) - \mathcal{I}(u_{\mathrm{in}})) - \epsilon_h^{\mathrm{LDA}}, \quad \epsilon_h^{\mathrm{LDA}} \geq 0 \qquad (85)$$

the first term representing the approximation of the net energy flux through the whole spatial domain. Estimate (85) proves the dissipative character of the scheme. However, in this case it is not clear how to simplify the first terms by means of the BCs, to obtain eventually a full proof of stability.

4.3.2 *Multidimensional upwind schemes: the N scheme*

The N scheme is perhaps the most successful first-order scheme for the solution of the advection equation. First proposed by Roe in the 1980s (Roe, 1987), it has been since then the basis for the construction of nonlinear positive and $\mathcal{LP}$ schemes. Thanks to its $\mathcal{MU}$ character, it has the lowest

numerical dissipation among first-order schemes (Paillère, 1995). It is defined by the following local nodal residuals:

$$\phi_i^{\mathrm{N}} = k_i^+(u_i - u_{\mathrm{in}}) \tag{86}$$

Being $\mathcal{MU}$, the N scheme differs from the LDA only in the 2-target case, in which a simple geometrical representation exists. In particular, we introduce the vectors $\vec{a}_1$ and $\vec{a}_2$, parallel to the edges $\overline{31}$ and $\overline{32}$ respectively, such that $\vec{a}_1 + \vec{a}_2 = \vec{a}$ (see Figure 6). Simple algebra shows that

$$\begin{aligned}\phi^E(\vec{a}) &= \int_E \vec{a}\cdot\nabla u_h \,\mathrm{d}x\,\mathrm{d}y \\ &= \phi^E(\vec{a}_1) + \phi^E(\vec{a}_2) = k_1(u_1 - u_3) + k_2(u_2 - u_3)\end{aligned}$$

which immediately gives for the N scheme

$$\begin{aligned}\phi_1^{\mathrm{N}} &= k_1(u_1 - u_3) = \phi^E(\vec{a}_1) \\ \phi_2^{\mathrm{N}} &= k_2(u_2 - u_3) = \phi^E(\vec{a}_2)\end{aligned}$$

In the 2-target case, the scheme reduces to first-order upwinding along the edges, with properly defined advection speeds.

N scheme: positivity and energy stability

To check that it verifies the local LED condition, we rewrite the N scheme as

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

Since $c_{ij}^E = -k_i^+ N k_j^- \geq 0$, the N scheme verifies Propositions 4, 5, and 6, and Theorem 2, and the related stability bounds. The time-step restrictions for its positivity and local positivity read

$$\Delta t \leq \frac{|S_i|}{(1-\theta)\sum_{E\in\mathcal{D}_i} k_i^+},\ \forall i \in \mathcal{T}_h$$

and

$$\Delta t \leq \frac{|E|}{3(1-\theta)k_i^+},\quad \forall E \in \mathcal{D}_i,\ \forall i \in \mathcal{T}_h \tag{87}$$

with $\theta \in [0, 1)$. These constraints can be shown to be larger than the corresponding ones of the upwind $\mathcal{FV}$-$\mathcal{RD}$ and Rv scheme (Paillère, 1995). In addition to this, we note that

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

which, as for 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 7 and 8. In particular, the energy evolution of the scheme can be easily shown to be given by

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{N}}}{\mathrm{d}t} = -\frac{1}{2}\oint_{\partial\Omega} \mathcal{I}_h \vec{a}\cdot\hat{n}\,\mathrm{d}l - \sum_{E\in\mathcal{T}_h}\sum_{i\in E}\sum_{j\in E} u_i \overline{M}_{ij}^{\mathrm{N}} u_j$$

where the boundary integral can be handled as done for the SUPG scheme, and $\overline{M}^{\mathrm{N}}$ is the (positive semidefinite) matrix energy operator (Abgrall and Barth, 2002; Barth, 1996; Abgrall and Mezine, 2003b)

$$\begin{aligned}\overline{M}^{\mathrm{N}} = &\frac{1}{2}\begin{bmatrix}k_1\\k_2\\k_3\end{bmatrix} N \begin{bmatrix}k_1\\k_2\\k_3\end{bmatrix}^T + \frac{1}{2}\begin{bmatrix}k_1^+ & 0 & 0\\ 0 & k_2^+ & 0\\ 0 & 0 & k_3^+\end{bmatrix} \\ &- \frac{1}{2}\begin{bmatrix}k_1^+\\k_2^+\\k_3^+\end{bmatrix} N \begin{bmatrix}k_1^+\\k_2^+\\k_3^+\end{bmatrix}^T + \frac{1}{2}\begin{bmatrix}-k_1^- & 0 & 0\\ 0 & -k_2^- & 0\\ 0 & 0 & -k_3^-\end{bmatrix} \\ &- \frac{1}{2}\begin{bmatrix}-k_1^-\\-k_2^-\\-k_3^-\end{bmatrix} N \begin{bmatrix}-k_1^-\\-k_2^-\\-k_3^-\end{bmatrix}^T\end{aligned} \tag{88}$$

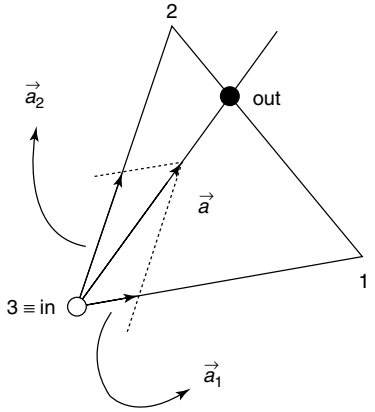

Figure 6. Geometry of $\mathcal{FS}$ schemes. N scheme in the two-target case.

4.3.3 *Relations between the N and LDA schemes: dissipation, nonhomogeneous problems*

Here, we elaborate on the relations between the N and the LDA schemes. We show that the N scheme can be written as the LDA scheme plus an anisotropic dissipation term. We start with the following observation. The definition of

the inflow state (78) is such that, for the N scheme, in the homogeneous case one has automatically

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

We can reverse things and, given ϕ^E, compute a state u_{in}^* from the satisfaction of the consistency constraint (13):

$$\sum_{j\in E} k_j^+(u_j - u_{\mathrm{in}}^*) = \phi^E \Longrightarrow u_{\mathrm{in}}^* = N\left(\sum_{j\in E} k_j^+ u_j - \phi^E\right) \tag{89}$$

Clearly, if ϕ^E is given by (52) (with $\mathcal{S} = 0$), using the relation $k_j = k_j^+ + k_j^-$, we get back $u_{\mathrm{in}}^* = u_{\mathrm{in}}$ as in (78). However, we can obtain additional information by using (89) in (86):

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

and finally

$$\phi_i^{\mathrm{N}} = \phi_i^{\mathrm{LDA}} + d_i^{\mathrm{N}}, \quad d_i^{\mathrm{N}} = k_i^+(u_i - u_{\mathrm{out}}) \tag{90}$$

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

$$\sum_{j\in E} d_j^{\mathrm{N}} = 0 \tag{91}$$

We can say more about this term by examining its contribution to the energy balance of the N scheme. Denoting the nodes of the element by $(1, 2, 3)$, we define the vector $d^{\mathrm{N}} = [d_1^{\mathrm{N}}, d_2^{\mathrm{N}}, d_3^{\mathrm{N}}]^T$ given by

$$\begin{aligned} d^{\mathrm{N}} &= D^{\mathrm{N}} U \\ D^{\mathrm{N}} &= \begin{bmatrix} k_1^+ & 0 & 0 \\ 0 & k_2^+ & 0 \\ 0 & 0 & k_3^+ \end{bmatrix} - \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix} N \begin{bmatrix} k_1^+ \\ k_2^+ \\ k_3^+ \end{bmatrix}^T \end{aligned} \tag{92}$$

with $U = [u_1, u_2, u_3]^T$. The matrix D^{N} is symmetric, and it is positive semidefinite, as shown by

$$\begin{aligned} \epsilon^{\mathrm{N}} = U^T D^{\mathrm{N}} U = {} & (u_1 - u_2) k_1^+ N k_2^+ (u_1 - u_2) \\ & + (u_1 - u_3) k_1^+ N k_3^+ (u_1 - u_3) \\ & + (u_2 - u_3) k_2^+ N k_3^+ (u_2 - u_3) \geq 0 \end{aligned} \tag{93}$$

Clearly d_i^{N} is a dissipation term. In particular, *the N scheme is more dissipative than the LDA scheme* (Abgrall, 2001; Abgrall and Mezine, 2003b):

$$\frac{\mathrm{d}\mathcal{E}_h^{\mathrm{N}}}{\mathrm{d}t} = \frac{\mathrm{d}\mathcal{E}_h^{\mathrm{LDA}}}{\mathrm{d}t} - \sum_{E\in\mathcal{T}_h} \epsilon^{\mathrm{N}} \leq \frac{\mathrm{d}\mathcal{E}_h^{\mathrm{LDA}}}{\mathrm{d}t} \tag{94}$$

Relations (90) and (91) give a simple means of extending the N scheme to more general situations. In particular, in the nonhomogeneous case $\mathcal{S} = \mathcal{S}(x, y)$, we have (using (52) and (90))

$$\phi_i^{\mathrm{N}} = \beta_i^{\mathrm{LDA}} \phi^E + d_i^{\mathrm{N}} = k_i^+(u_i - u_{\mathrm{in}}) - \sum_{j\in E} \frac{|E|}{3} \beta_i^{\mathrm{LDA}} \mathcal{S}_j \tag{95}$$

with u_{in} as in (78). One can show that, if the source term is independent of the solution, scheme (95) verifies a modified discrete maximum principle (see Ricchiuto, 2005; Sidilkover and Roe, 1995 for more details). Scheme (95) was initially proposed in Sidilkover and Roe (1995).

4.4 Nonlinear $\mathcal{RD}$ schemes

Nonlinear schemes are needed to combine linearity preservation and LED. The interest in $\mathcal{FS}$ discretizations is largely due to the success of the nonlinear PSI scheme of Struijs, Struijs, Deconinck and Roe (1991). For steady scalar advection, the PSI scheme has been proved to perform better than standard second-order limited $\mathcal{FV}$ schemes, especially on irregular grids (Struijs, Deconinck and Roe, 1991; Paillère, 1995; Roe and Sidilkover, 1992; Sidilkover and Roe, 1995; Abgrall and Mezine, 2004; Abgrall and Mezine, 2003b). Being completely parameter free, it is an interesting alternative to $\mathcal{FE}$ schemes with shock-capturing terms (Paillère, 1995; Carette *et al.*, 1995). Unfortunately, when dealing with inhomogeneous or time-dependent problems and systems, 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 references given in the introduction.

Here we consider two 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 that use as linear LED scheme the N scheme.

4.4.1 Blended schemes

Given a $\mathcal{LP}$ scheme defined by the split residuals $\phi_i^{\mathcal{LP}}$, and a linear LED first-order scheme, defined by the local nodal residuals ϕ_i^{LED}, a blended scheme is defined by

$$\phi_i = (1 - \Theta(u_h))\phi_i^{\mathcal{LP}} + \Theta(u_h)\phi_i^{\mathrm{LED}} \tag{96}$$

where $\Theta(u_h)$ is a blending parameter, which must ensure that $\phi_i = \mathcal{O}(h^3)$ in smooth regions, and 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. When blending the LDA and the N schemes, the blending approach has an interesting interpretation. In particular, using (90) we can write that

$$\phi_i = (1 - \Theta(u_h))\phi_i^{\mathrm{LDA}} + \Theta(u_h)\phi_i^{\mathrm{N}} = \phi_i^{\mathrm{LDA}} + \Theta(u_h)d_i^{\mathrm{N}} \tag{97}$$

Blending the LDA and the N scheme is equivalent to adding a nonlinear dissipation term to the LDA scheme. Defining Θ in a very rigorous way might not be extremely important in practice, as shown by the fact that the *heuristic* definition of the blending parameter of Deconinck and collaborators Sermeus and Deconinck (2005) and Deconinck, Sermeus and Abgrall (2000).

$$\Theta(u_h) = \frac{|\phi^E|}{\sum_{j\in E}|\phi_j^{\mathrm{N}}|} \in [0, 1] \tag{98}$$

has given good results in several applications (Sermeus and Deconinck, 2005; Henriques and Gato, 2004; Abgrall and Mezine, 2003b; Csík, Deconinck and Poedts, 2001; Csík, Ricchiuto and Deconinck, 2003b). A rigorous study of this problem is found in Abgrall (2001) and Abgrall and Mezine (2003b). 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 $\Theta(u_h)$.

4.4.2 Limited nonlinear schemes

Several generalizations of the PSI scheme of Struijs exist (see e.g. Paillère, 1995 for a discussion). The most general formulation is obtained by introducing the framework of the so-called *limited* schemes (Paillère, 1995; Abgrall and Mezine, 2003b; Abgrall and Mezine, 2004; Abgrall and Roe, 2003). Consider a first-order linear $\mathcal{FS}$ scheme, with split residuals ϕ_i^{LED}, verifying the subelement LED condition. Suppose that we have a *continuous nonlinear mapping* $\varphi(x_0, x_1, x_2, x_3) : \mathbb{R}^4 \mapsto \mathbb{R}^3$ such that

$$\varphi(x_0, x_1, x_2, x_3) = (x_0\, y_1,\ x_0\, y_2,\ x_0\, y_3) \tag{99}$$

with

$$x_j = 0 \Longrightarrow\ y_j = 0 \quad \forall j = 1, 2, 3 \tag{100}$$

$$x_j \cdot (x_0 y_j) \geq 0 \quad \forall j = 1, 2, 3 \tag{101}$$

$$|y_j| < \infty \quad \forall j = 1, 2, 3 \tag{102}$$

$$y_1 + y_2 + y_3 = 1 \tag{103}$$

A limited $\mathcal{FS}$ scheme is obtained as

$$(\phi_1, \phi_2, \phi_3) = \varphi(\phi^E, \phi_1^{\mathrm{LED}}, \phi_2^{\mathrm{LED}}, \phi_3^{\mathrm{LED}}) \tag{104}$$

The properties of such a scheme are determined by those of the mapping. In particular, (103) guarantees that the scheme verifies the consistency condition (13). Property (102), together with (99), and with the continuity of the mapping, guarantees that the scheme is $\mathcal{LP}$. Moreover, conditions (100) and (101) guarantee that, if $\phi^E \neq 0$, then if $\phi_j^{\mathrm{LED}} = 0$ also $\phi_j = 0$, otherwise one has

$$\phi_j = x_0 y_j = \frac{x_0 y_j}{x_j} x_j = \alpha_j x_j = \alpha_j \phi_j^{\mathrm{LED}}$$

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

Hence the resulting scheme also verifies the subelement LED condition. There are quite a number of constructions leading to functions φ verifying (99)–(103). A review can be found in Abgrall and Mezine (2003b), Abgrall and Mezine (2004), and Abgrall and Roe (2003). In particular, starting from the N scheme, one obtains the PSI scheme of Struijs with the choice

$$\varphi(x_0, x_1, x_2, x_3) = \frac{1}{\sum_{j=1,3}(x_0 x_j)^+}\left((x_0 x_1)^+, (x_0 x_2)^+, (x_0 x_3)^+\right) x_0 \tag{105}$$

This formulation of the PSI scheme has been known since long. However, only lately this more general framework has emerged as a way of constructing nonlinear schemes for time-dependent problems and systems (Paillère, 1995; Abgrall and Mezine, 2004). We remark that (105) can be rewritten in the simpler form

$$\beta_i = \frac{\max(0, \beta_i^{\mathrm{N}})}{\sum_{j=1,3}\max(0, \beta_j^{\mathrm{N}})}, \quad \beta_j^{\mathrm{N}} = \frac{\phi_j^{\mathrm{N}}}{\phi^E} \tag{106}$$

which is how the limited N (LN) scheme is normally presented in literature (Paillère, 1995; Abgrall and Mezine, 2004; Abgrall and Roe, 2003). 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. We recall once more that, for steady advection, in Abgrall (2001) it has been shown that the scheme obtained by applying (105) to the N scheme can be written as a blended LDA/N scheme. Generally speaking, often the limited schemes work quite well even when blended schemes fail (Csík, Ricchiuto and Deconinck, 2002; Abgrall and Mezine, 2003b; Abgrall and Mezine, 2004). Nevertheless, mapping (105), which is the one commonly used in practice, is known since a very long

time and improved constructions still have to appear. The study and the understanding of these nonlinear mappings is one of the most important subjects of future research. We mention, in this regard, the recent work of Abgrall (2006) and Ricchiuto and Abgrall (2006) investigating the algebraic well-posedness of limited schemes. The well-posedness of the construction has been studied in Ricchiuto, Csík and Deconinck (2005), where the following simple result has been proved.

Proposition 10 (Well-posedness of the mapping — sufficient condition) *Given a linear scheme satisfying the subelement LED condition, defined by the split residuals* ϕ_j^{LED}*, a condition to construct a well-posed nonlinear mapping satisfying properties* (99)–(103) *is that*

$$\phi^E \sum_{j\in E} \phi_j^{\mathrm{LED}} > 0 \tag{107}$$

Even though the last condition seems trivial, schemes violating (107) have been considered in some works (see e.g. Abgrall and Roe, 2003; Ricchiuto, Csík and Deconinck, 2005; Ricchiuto, 2005 for a discussion).

Nonlinear schemes: energy stability

This is an *ongoing* research topic. Only some qualitative arguments can be given. We only consider the blended LDA/N scheme and the limited schemes. For the former, using the result of the analysis of the LDA and N schemes, the energy evolution equation can be easily shown to be

$$\begin{aligned}\frac{d\mathcal{E}_h}{\mathrm{d}t} &= -\sum_{E\in\mathcal{T}_h}\left(\sum_{j\in E} k_j^+\right)\left(\mathcal{I}(u_{\mathrm{out}}) - \mathcal{I}(u_{\mathrm{in}})\right)\\ &- \epsilon_h^{\mathrm{LDA}} - \sum_{E\in\mathcal{T}_h}\Theta(u_h)\epsilon^{\mathrm{N}} \quad \epsilon_h^{\mathrm{LDA}}, \epsilon^{\mathrm{N}}, \Theta(u_h) \geq 0\end{aligned} \tag{108}$$

having used (85), and with ϵ^{N} given by (93). The last expression clearly shows the dissipative character of the blended LDA/N scheme, due to its $\mathcal{MU}$ character. The last expression also applies to the PSI scheme of Struijs (LN scheme, obtained with (105)), which, for a particular choice of $\Theta(u_h)$, has been proved to reduce to a blended LDA/N scheme in Abgrall (2001). So far, $\mathcal{MU}$ seems enough to guarantee good stability properties. Unfortunately, similar stability properties cannot be shown for nonlinear schemes obtained by applying, for example, mapping (105) to non-$\mathcal{MU}$ schemes, such as the Rv or the $\mathcal{FV}$-$\mathcal{RD}$ schemes. Even in the case of the PSI scheme of Struijs, a local analysis shows that in the 2-target cases local sources of energy instability might appear (see Barth, 1996 and also Ricchiuto, 2005), even though in practice the scheme is perfectly stable, confirming the validity of (108). Conversely, on several occasions these instabilities have been shown to pollute the numerical results, when applying the limiting technique to non-$\mathcal{MU}$ schemes. The symptoms of this lack of stability are poor iterative and grid convergence (Abgrall and Mezine, 2003b; Ricchiuto, 2005) (only first order, even if the schemes are $\mathcal{LP}$ by construction). We limit ourselves to the observation that the limiting approach is built entirely on stability considerations in the maximum (L^∞) norm, and it does not take into account in any way either the energy (L^2) norm, or the directional propagation of the information typical of hyperbolic PDEs. In this respect, nonlinear limited $\mathcal{RD}$ schemes are substantially different from stabilized Galerkin $\mathcal{FE}$ schemes with nonlinear shock-capturing ($\mathcal{SC}$), which have by construction a dissipative character. The energy stability of the resulting schemes is quite clear (Barth, 1998). However, L^∞ stability is only recovered indirectly for nonlinear $\mathcal{FE}$ schemes, thanks to the regularization of the solution introduced by the additional nonlinear dissipation (Szepessy, 1989). 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. Again we refer to Abgrall (2006) and Ricchiuto and Abgrall (2006) for a recent analysis of the problem. The study of improved and more general constructions certainly deserves more attention.

4.5 Nonlinear problems

We now discuss some issues related to the extension of $\mathcal{RD}$ schemes to the case of fully nonlinear conservation laws such as (4). We mainly consider the issues of constructing conservative and stable discretizations. As far as accuracy is concerned, the analysis of Section 3.2 applies equally to the nonlinear case.

4.5.1 Conservation

In this and in the next paragraph, we consider the homogeneous counterpart of (4), which reads

$$\frac{\partial u}{\partial t} + \nabla\cdot\mathcal{F}(u) = 0 \quad \text{on} \quad \Omega\subset\mathbb{R}^2 \tag{109}$$

or, in quasi-linear form

$$\frac{\partial u}{\partial t} + \vec{a}(u)\cdot\nabla u = 0, \quad \vec{a}(u) = \frac{\partial \mathcal{F}(u)}{\partial u} \tag{110}$$

The schemes presented in the previous section rely on the use of the quasi-linear form (110). However, even for smooth initial and boundary data, nonlinear problems evolve discontinuous solutions across which the relevant form of the problem is obtained by integrating (109) in space-time (Serre, 1999; Evans, 1998), while (110) cannot be used unless appropriate linearizations are introduced. As a motivational example, consider (109) with the *exponential* flux

$$\mathcal{F}(u) = (e^u, u)$$

We take $\Omega = [-0.025, 1.2] \times [0, 0.5]$ with BCs:

$$u(x, y = 0) = \begin{cases} \sin(2\pi x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}, \quad u(-0.025, y) = 0 \quad (111)$$

On a fine unstructured discretization of Ω ($h = 1/200$), we compute the steady solution of this problem with the LN scheme obtained by blindly linearizing (110) on each element and applying mapping (106) to the N scheme (86), and with the same scheme but with a more accurate mean-value linearization. In other words, on each element we solve the linearized problem

$$\frac{\partial u}{\partial t} + \tilde{a} \cdot \nabla u = 0$$

with $\tilde{a}$ given by

$$\tilde{a} = \frac{1}{3} \sum_{j \in E} \vec{a}(u_j)$$

for the first scheme (referred to as *nonconservative* limited N scheme (LN-NC)), and with $\tilde{a}$ obtained evaluating

$$\tilde{a} = \frac{1}{|E|} \int_E \vec{a}(u_h) \, \mathrm{d}x \, \mathrm{d}y = \frac{1}{|E|} \int_E (e^{u_h}, 1) \, \mathrm{d}x \, \mathrm{d}y \quad (112)$$

with a four-points Gaussian formula for the second scheme (referred to as *conservative* limited N scheme (LN-C)). Contour plots of the solutions obtained with the two schemes are reported on Figure 7. From the plots, we see that, even if the boundary data are continuous, with piecewise continuous derivatives, the solution contains a *shock* that develops at a finite and relatively small distance from the lower boundary, where the smooth data are imposed. At first sight, the two solutions look identical. However, a closer examination shows some major differences in the approximation of the discontinuity. This is shown in Figure 8(a), where we have reported a line plot of the two solutions at $y = 0.5$ (upper boundary), and in Figure 8(b) a close-up view of the solution of the LN-NC scheme, superimposing the direction of the shock, as computed by the LN-C scheme.

The two schemes give a different prediction of angle and position of the shock. An explanation of this fact is the following. Suppose that the error made when approximating (112) with four Gaussian points is small enough, in particular, that we can assume that for the LN-C scheme equation (112) is integrated exactly and so $\tilde{a}$ is an exact mean-value linearization. In this case, we have, using (52) and applying Gauss' theorem,

$$\phi^E = \sum_{j \in E} \tilde{k}_j u_j = \int_E \tilde{a} \cdot \nabla u_h \, \mathrm{d}x \, \mathrm{d}y = \int_E a(u_h) \cdot \nabla u_h \, \mathrm{d}x \, \mathrm{d}y$$
$$= \int_E \nabla \cdot \mathcal{F}(u_h) \, \mathrm{d}x \, \mathrm{d}y = \oint_{\partial E} \mathcal{F}(u_h) \cdot \hat{n} \, \mathrm{d}l \quad (113)$$

Hence, if the error made in the evaluation of (112) is negligible, the LN-C scheme is consistent with the integral form of (109), thus giving a correct approximation of the discontinuity. The same cannot be said for the LN-NC scheme, for which the third equality in (113) is not true. This shows that, in extending $\mathcal{RD}$ schemes to nonlinear conservation laws, care must be taken in ensuring that the element residual is a consistent approximation of the flux

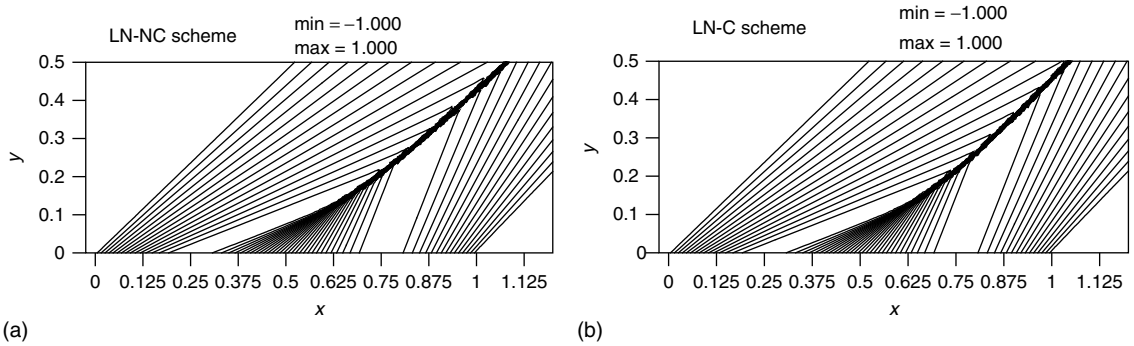

Figure 7. Nonlinear problem with *exponential flux*. Contour plot of the solution obtained with the LN-NC (a) and LN-C (b) schemes.

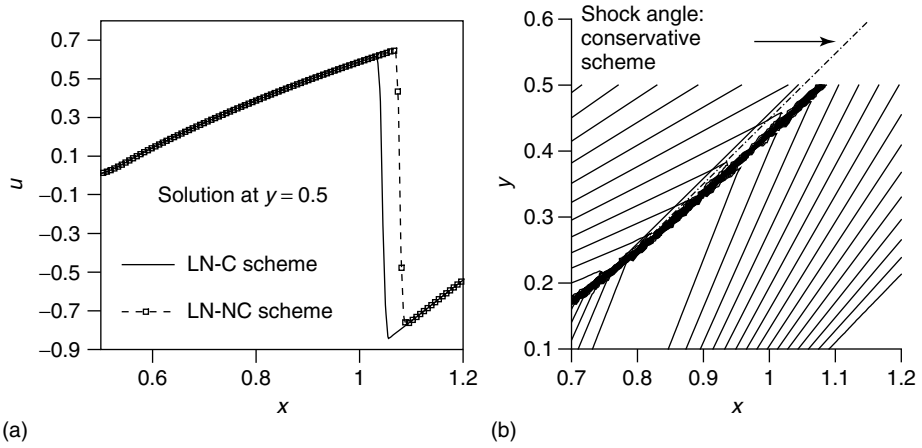

Figure 8. Nonlinear problem with *exponential flux*: conservation error. a: solution at $y = 0.5$ in vicinity of the shock. b: close-up of the shock, solution of the LN-NC with conservative shock angle superimposed.

balance over the element, before the distribution step. This leads to the following definition.

Definition 9 (Conservative $\mathcal{RD}$ scheme) *A $\mathcal{RD}$ scheme is conservative if there existe a continuous approximation of the flux $\mathcal{F}_h$ such that*

$$\phi^E = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n}\, \mathrm{d}l \tag{114}$$

In Abgrall, Mer and Nkonga (2002), 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 LW theorem. Conservative schemes guarantee a correct approximation of the integral form of (109), hence yielding a correct prediction of steady discontinuities. Unfortunately, we have arrived at a problem of *incompatibility* between the use of the integral form of equation (109), needed to guarantee the approximation of the correct weak solution, and the use of the flux Jacobians, needed in the definition of the k_j parameters used in the distribution of the residual. A discussion of this problem and two alternative possible solutions can be found in Abgrall and Barth (2002) and Csík, Ricchiuto and Deconinck (2002), and is reviewed in the following paragraphs.

4.5.2 *Conservative $\mathcal{RD}$: accurate quadrature of the quasi-linear form*

The analysis of our motivational example leads to the approach used in Abgrall and Barth (2002) to construct $\mathcal{FS}$ schemes based on the use of the quasi-linear form (110), still guaranteeing a correct approximation of weak discontinuous solutions. The basic idea is contained in equation (113): if $\widetilde{a}$ is computed exactly, the schemes obtained in this way obey Definition 9 with $\mathcal{F}_h = \mathcal{F}(u_h)$, and u_h as in (9). However, the derivation of such an *exact mean-value linearization* of the flux Jacobian $\vec{a}(u)$ can be difficult, and in the case of a system even impossible. This was the motivation to introduce in Abgrall and Barth (2002) an *approximate mean-value linearization* obtained with the Gaussian integration

$$\overline{a} = |E| \sum_{l=1}^{\mathrm{N_Q}} \omega_l \vec{a}(u(x_l, y_l)), \quad (x_l, y_l) \in E \tag{115}$$

where ω_l is the quadrature weight corresponding to the lth Gaussian point (x_l, y_l). This leads to

$$\phi^E = \sum_{j \in E} \overline{k}_j u_j = \int_E \nabla \cdot \mathcal{F}(u_h)\, \mathrm{d}x\, \mathrm{d}y + \mathrm{R}_{\mathrm{N_Q}} \tag{116}$$

where $\mathrm{R}_{\mathrm{N_Q}}$ is the *conservation error* due to the approximate integration of $\vec{a}(u_h)$. The properties of the Gaussian integration, namely the behavior of the quadrature error, allows the authors of Abgrall and Barth (2002) to prove that

1. provided that the number of quadrature points $\mathrm{N_Q}$ is large enough, the conservation error due to the approximate integration is strictly smaller than the discretization error of the schemes;
2. LW theorem: provided that the number of quadrature points $\mathrm{N_Q}$ 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*.

This approach indeed represents a solution for the extension of $\mathcal{RD}$ schemes to general nonlinear conservation laws. It is mathematically sound, and it allows us to apply the maximum discrete principle analysis of Section 3.3 also in the

nonlinear case. However, it has the drawback of requiring the evaluation of the flux Jacobians in several quadrature points, which becomes computationally demanding when approximating solutions to systems, especially in the presence of strong discontinuities (see Abgrall and Barth, 2002 for more).

4.5.3 *Contour integration of the fluxes and monotone schemes*

A simpler, yet very effective, alternative approach is proposed in Csík, Ricchiuto and Deconinck (2002). The first element of the construction is the definition of the element residual. Given a continuous approximation of the flux $\mathcal{F}_h$, one computes ϕ^E as

$$\phi^E = \oint_{\partial E} \mathcal{F}_h \cdot \hat{n}\, \mathrm{d}l = \sum_{l_j=1}^{3} \mathcal{F}^{l_j} \cdot \vec{n}_{l_j}$$

$$\text{with} \quad \mathcal{F}^{l_j} = \sum_{p=1}^{\mathrm{N_C}} \omega_p \mathcal{F}_h(x_p, y_p), \quad (x_p, y_p) \in l_j \qquad (117)$$

l_j being the jth edge of E, $\vec{n}_{l_j}$ is its exterior normal, scaled by its length, and ω_p is the weight of the pth quadrature point on l_j. As before, the computation of the residual is based on a quadrature formula, but however, now Definition (117) satisfies by construction (114). Hence, conservation is in this case guaranteed by construction. However, we still need to specify how the flux Jacobians can be used to distribute ϕ^E. We will distinguish between the case of a $\mathcal{LP}$ scheme and the one of schemes that are positive when applied to a linear problem.

$\mathcal{LP}$ schemes The case of $\mathcal{LP}$ schemes is quite simple. These schemes are defined by

$$\phi_i = \beta_i \phi^E$$

with β_i uniformly bounded and respecting, by construction, the consistency relation

$$\sum_{j \in E} \beta_j = 1$$

The dependence of the distribution coefficients on the k_j parameters does not alter any of these two properties (boundedness and consistency): we can use for the computation of the β_i's the parameters

$$k_j = \frac{\vec{a}(u_E) \cdot \vec{n}_j}{2} \qquad (118)$$

with u_E an arbitrary average of u_h over E.

Positive schemes The case of the positive schemes is more difficult. This is easily seen for the N scheme, whose definition (equation (86)) is entirely based on the quasi-linear form of the problem. As proposed in Csík, Ricchiuto and Deconinck (2002), the solution of this problem is difficult to put in a general framework. We use instead the formulation proposed in Ricchiuto (2005), where it has been underlined as to how the conservative N scheme of Csík, Ricchiuto and Deconinck (2002) is a particular case of a class of positive $\mathcal{RD}$ schemes, which can be written as

$$\phi_i = \beta_i \phi^E + d_i = \beta_i \phi^E + \sum_{j \in E} D_{ij}(u_i - u_j), \quad D_{ij} \geq 0 \qquad (119)$$

with bounded distribution coefficients β_i and dissipation terms d_i respecting the consistency relations

$$\sum_{j \in E} \beta_j = 1, \quad \sum_{j \in E} d_j = 0$$

Owing to the last relations, *independently of the linearization used to evaluate* β_i *and* d_i, conservative variants of the positive schemes are obtained just by using in (119) the residual computed according to (117). The cases of the N scheme and of the Rv scheme are easily obtained from equations (90) and (71), giving

$$\phi_i^{\text{N-C}} = \beta_i^{\text{LDA}} \phi^E + d_i^{\text{N}} = \beta_i^{\text{LDA}} \phi^E + \sum_{j \in E} k_i^+ N k_j^+ (u_i - u_j) \qquad (120)$$

for the N scheme (where C stands for conservative), and giving for the Rv scheme

$$\phi_i^{\text{Rv-C}} = \frac{1}{3}\phi^E + d_i^{\text{Rv}} = \frac{1}{3}\phi^E + \frac{1}{3}\alpha \sum_{\substack{j \in E \\ j \neq i}} (u_i - u_j) \qquad (121)$$

We refer the reader to Ricchiuto (2005) for a discussion on the relations of this approach with other conservative formulations proposed in literature. Relations with $\mathcal{FV}$ schemes, in particular with the 1D scheme of Huang (1981), are discussed in Ricchiuto *et al.* (2003). Note that conservative positive schemes trivially verify condition (107), since for $\phi^E \neq 0$

$$\phi^E \sum_{j \in E} \phi_j = \left(\phi^E\right)^2 > 0$$

making the constructions of limited nonlinear $\mathcal{LP}$ positive schemes always well-posed.

4.5.4 *Positive schemes for nonlinear problems*

The fact that nonlinear problems admit discontinuous solutions makes the need for schemes satisfying a discrete

maximum principle even greater. For $\mathcal{RD}$ schemes, we have to distinguish whether one makes use of the approach of Abgrall and Barth (2002), based on the Gaussian quadrature of the quasi-linear form, or of the conservative formulation of Csík, Ricchiuto and Deconinck (2002), based on boundary integration of the fluxes. As already remarked, in the first case the use of the quasi-linear form allows to apply all the results of Section 3.3 to the nonlinear case. We shall then focus on the second approach, for which no results are available in the published literature, with the exception of Ricchiuto (2005). Even though the schemes of Csík, Ricchiuto and Deconinck (2002) are based on the use of the (nonlinear) fluxes for the definition of the local residual, for the analysis one can anyway make use of the quasi-linear form. In particular, we make the following assumption.

Assumption 1. Given a $\mathrm{N_c}$-points line quadrature formula used to evaluate (117), it is possible to find a $\mathrm{N_q}$-surface quadrature rule to be used in (115), such that the equivalence

$$\begin{aligned}\phi^E &= \sum_{l_j=1}^{3}\sum_{p=1}^{\mathrm{N_c}} \omega_p \mathcal{F}(u_{p,l_j})\cdot\vec{n}_{l_j} \\ &= |E|\sum_{l=1}^{\mathrm{N_q}} \omega_l \vec{a}(u_l)\cdot\nabla u_h\big|_E = \sum_{j\in E}\overline{k}_j u_j \end{aligned} \tag{122}$$

holds up to the smallest between the quadrature error in (117), and the one in (116).

In the following, we shall use the notation $\overline{k}_j$, to denote the scalar upwind parameters based on the approximate mean-value linearization, while using k_j to denote the ones based on any (also inexact) arbitrary linearization. Using the general representation of a monotone $\mathcal{RD}$ (119), we have to then to analyze schemes of the form

$$\phi_i = \beta_i \sum_{j\in E}\overline{k}_j u_j + \sum_{j\in E} D_{ij}(u_i - u_j) \tag{123}$$

where in general D_{ij} is evaluated making use of the k_j's. Now we can recast our prototype in the form (30), with

$$c_{ii} = \beta_i\overline{k}_i + \sum_{j\in E} D_{ij}, \quad c_{ij} = \beta_i\overline{k}_j - D_{ij}\ ; \quad D_{ij}\geq 0$$

This notation allows to prove two results, one positive and the other (unfortunately) negative.

Proposition 11 (Rv-C (Rusanov scheme based on Contour integration) scheme and subelement LED) *The Rv-C scheme* (121) *respects the subelement LED condition, provided that* α *in* (121) *is chosen big enough.*

Proof. Trivially, for α big enough $\phi_i^{\mathrm{Rv}-C} = \sum_{\substack{j\in E\\ j\neq i}} c_{ij}(u_i - u_j)$, with $c_{ij} = \frac{1}{3}(\alpha - \overline{k}_j) \geq 0$. □

The last proposition makes the Rv-C scheme a very good candidate to be used as a basis for the construction of a positive limited nonlinear $\mathcal{LP}$ scheme. Unfortunately, as underlined in Section 4.4, the limited knowledge on the (L^2) stability of the limited schemes does not guarantee that the resulting scheme would have the expected convergence properties. Even though recent developments (see Abgrall, 2006; Ricchiuto and Abgrall, 2006) would allow for stable constructions based on centered low-order schemes, the basic available technology works best with schemes with a pronounced upwind character, such as the N scheme for which, unfortunately, we have the following negative result.

Proposition 12 (N-C (Narrow scheme based on Contour integration) scheme and subelement LED) *The N-C scheme* (120) *cannot be proven to respect the subelement LED condition. In particular, the scheme is prone to the violation of this condition in multiple-target elements.*

Proof. We start by writing (123) for the N-C scheme:

$$\phi_i^{\mathrm{N}-C} = k_i^+ N\sum_{j\in E}\overline{k}_j u_j + \sum_{\substack{j\in E\\ j\neq i}} k_i^+ N k_j^+ (u_i - u_j)$$

Since the $\overline{k}_j$s sum up to zero over an element (see (54)), one can show that

$$\begin{aligned}\phi_i^{\mathrm{N}-C} &= \sum_{\substack{j\in E\\ j\neq i}} k_i^+ N\overline{k}_j(u_j - u_i) + \sum_{\substack{j\in E\\ j\neq i}} k_i^+ N k_j^+ (u_i - u_j) \\ &= \sum_{\substack{j\in E\\ j\neq i}} k_i^+ N(k_j^+ - \overline{k}_j^+)(u_i - u_j) - \sum_{\substack{j\in E\\ j\neq i}} k_i^+ N\overline{k}_j^-(u_i - u_j)\end{aligned}$$

If $k_j = \overline{k}_j$, as in the linear case or when using the formulation of Abgrall and Barth (2002), the scheme reduces to its standard expression, with $c_{ij} = -\overline{k}_i^+ N\overline{k}_j^- \geq 0$, proving the local LED condition in the nonlinear case. In general, however

$$c_{ij} = k_i^+ N(k_j^+ - \overline{k}_j^+) - k_i^+ N\overline{k}_j^-, \quad -k_i^+ N\overline{k}_j^- \geq 0$$

Since the sign of the first term on the right-hand side is unknown, we cannot prove the subelement LED condition. Consider now the multiple-target situation in which $k_i, \overline{k}_i, \overline{k}_j > 0$ for some $j \neq i$:

$$c_{ij} = k_i^+ N(k_j^+ - \overline{k}_j^+)$$

where the beneficial effect of the second term has disappeared. The sign of c_{ij} could be either positive or negative, depending on the local structure of the solution and 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. □

This result seems to spoil the hopes of constructing a nonoscillatory second-order nonlinear scheme, based on the use of the N-C scheme. On two-dimensional triangular grids the non-LED character of the N-C scheme could be limited to two-target elements, but in three dimensions things could get worse, owing to the presence of a larger number of two-target tetrahedra and of three-target ones. In practice, these effects have *never been observed* in any numerical result, in two and three space dimensions, for scalar problems and for systems (Csík, Ricchiuto and Deconinck, 2002; Ricchiuto, Csík and Deconinck, 2005; Ricchiuto, 2005; Ricchiuto, Abgrall and Deconinck, 2007). Extensions of the N-C scheme to meshes composed of quadrilaterals (Quintino *et al.*, 2002; De Palma *et al.*, 2006; Abgrall and Marpeau, 2007) have also been proven to yield nonoscillatory numerical solutions. We believe that the monotone resolution of discontinuities observed in practice is due partly to a compensation of the violation of the local LED condition when assembling the contributions of all the elements, and partly to the dissipative character of the scheme, which might be enough to dissipate weak new local extrema, eventually appearing in the numerical solution.

4.5.5 *A note on stability: conservative $\mathcal{RD}$ and entropy*

In the nonlinear case, the L^2 norm (energy) stability analysis is replaced by a better suited tool: the entropy stability analysis. The study of the stability of $\mathcal{RD}$ schemes in the so-called *entropy norm* is formally very difficult. Very few results are available in the published literature (see Abgrall and Mezine, 2003b for a review). For this reason this subject is left out of the paper. The reader is referred to Abgrall and Barth (2001), Abgrall and Barth (2002), Abgrall and Mezine (2003b) and Ricchiuto (2005) and references therein for further information.

5 EXTENSION TO TIME-DEPENDENT PROBLEMS

5.1 Preliminaries

This section considers the extension of $\mathcal{RD}$ schemes to the approximation of solutions to (4) in the time-dependent case. Common experience is that the prototype scheme of Definition 1 yields in practice first order of accuracy in unsteady computations, whatever be the distribution strategy adopted (Struijs, 1994; Maerz and Degrez, 1996; Ferrante and Deconinck, 1997). There is perhaps only one exception to this, represented by a LW discretization, which we shall discuss later.

In general, however, it seems that achieving second (or higher) order of accuracy in time-dependent computations requires the time derivative to be consistently introduced in the element residual. A heuristic explanation can be obtained as follows. First, redefine the source term as

$$\widetilde{\mathcal{S}}(u, x, y, t) = -\frac{\partial u}{\partial t} + \mathcal{S}(u, x, y, t)$$

Then, repeat the analysis of Section 3.2 for the (pseudo-) steady problem

$$\nabla \cdot \mathcal{F}(u) = \widetilde{\mathcal{S}}(u, x, y, t)$$

Proceeding exactly as in Section 3.2, we can write down the truncation error *due to the spatial approximation*

$$\begin{aligned} TE(w_h) &= \int_\Omega \varphi_h \big(-\widetilde{\mathcal{S}}_h(w_h, x, y, t) + \nabla \cdot \mathcal{F}_h(w_h)\big)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad + \frac{1}{K} \sum_{E \in \mathcal{T}_h} \sum_{i \in E} \sum_{j \in E} (\varphi_i - \varphi_j)\big(\phi_i^E(w_h) - \phi_i^{\mathrm{G}}(w_h)\big) \\ &= \underbrace{\int_\Omega \varphi_h \left(\frac{\partial w_h}{\partial t} + \nabla \cdot \mathcal{F}_h(w_h) - \mathcal{S}_h(w_h, x, y, t) \right) \mathrm{d}x\, \mathrm{d}y}_{I} \\ &\quad + \underbrace{\Delta TE(w_h)}_{II} \end{aligned} \tag{124}$$

for a given smooth exact solution of the time-dependent problem w, and C^1-class function φ with compact support. We recall that in equation (124) the term I is associated to the error introduced by the choice of the discrete polynomial approximation of the unknown, the flux, and the source term, while the second term represents the additional error introduced by the $\mathcal{RD}$ discretization.

From (124), the analysis proceeds exactly as in Section 3.2. In particular, kth order schemes must verify (25), and $\mathcal{LP}$ schemes (cf. Definition 3) are formally high order. More importantly, following the analysis of source term discretizations at the end of Section 3.2, we conclude that, since pointwise discretizations of the source term $\widetilde{\mathcal{S}}$ are generally first-order accurate in space, the prototype $\mathcal{RD}$ scheme of Definition 1 will be in general only first order during the transient. This is true *no matter what the approximation of the time derivative is, since the lack of accuracy is due to an inconsistency in*

the spatial discretization. Concerning the technical details of the analysis, no major differences are present with respect to what we have seen in Section 3.2. The reader is referred to Ricchiuto, Abgrall and Deconinck (2007) and Rossiello *et al.* (2007) for a detailed study, also including the influence of the choice of the discretization of the time derivative.

In order to contruct higher-order schemes for time-dependent problems, in the following section we introduce a more general prototype. Even though this can be done in a very general fashion, for simplicity we focus on a particular case of second-order discretizations. Additional references are given in the text, allowing the reader to have a wider overview on the subject.

5.2 A more general prototype

We assume to be given the set of nodal values of u at time t^n, $\{u_i^n\}_{i\in\mathcal{T}_h}$. Next, we note that in the space-time slab $\Omega \times [t^n,\, t^{n+1}]$, each element E in the mesh defines a prism in space-time, defined as (see Figure 9)

$$P_E^{n+1/2} := E \times [t^n,\, t^{n+1}] \tag{125}$$

By abuse of notation, we shall say that $P_E^{n+1/2} \in \mathcal{D}_i$ if $E \in \mathcal{D}_i$. In addition, we denote by u^n and u^{n+1} the piecewise linear discrete approximations (cf. equation (9))

$$u^n = \sum_{i\in\mathcal{T}_h} \psi_i(x,\, y) u_i^n\,, \quad u^{n+1} = \sum_{i\in\mathcal{T}_h} \psi_i(x,\, y) u_i^{n+1}$$

with $\{\psi_i\}_{i\in\mathcal{T}_h}$ being the piecewise linear finite-element basis functions verifying (11). With this notation, we give the following characterization in the scalar case.

Definition 10 (Space-time $\mathcal{RD}$) *A space-time $\mathcal{RD}$ or space-time $\mathcal{FS}$ scheme is defined as one that, given u^n, the discrete approximation in space of u at time t^n, and given a continuous discrete representation in space and time of the unknown u, denoted by u_h, and of the flux and of the source term, $\mathcal{F}_h$ and $\mathcal{S}_h$ respectively, computes the unknowns $\{u_i^{n+1}\}_{i\in\mathcal{T}_h}$ as follows:*

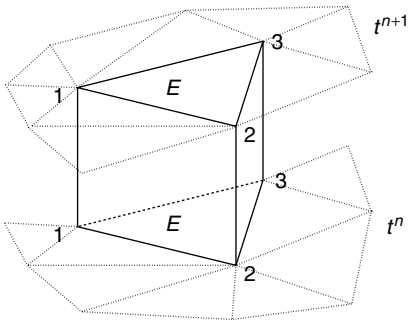

Figure 9. Space-time prism $P_E^{n+1/2} = E \times [t^n, t^{n+1}]$.

1. $\forall E \in \mathcal{T}_h$ *compute the space-time residual*

$$\begin{aligned}\Phi^{P_E^{n+1/2}} &= \int_{P_E^{n+1/2}} \left(\frac{\partial u_h}{\partial t} + \nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\ &= \int_E \int_{t^n}^{t^{n+1}} \left(\frac{\partial u_h}{\partial t} + \nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\end{aligned} \tag{126}$$

2. $\forall E \in \mathcal{T}_h$ *distribute fractions of $\Phi^{P_E^{n+1/2}}$ to the nodes of E. Denoting by $\Phi_i^{P_E^{n+1/2}}$ the split residual or local nodal residual for node $i \in E$, one must have by construction*

$$\begin{aligned}\sum_{j\in E} \Phi_j^{P_E^{n+1/2}} &= \Phi^{P_E^{n+1/2}} \\ &= \int_{P_E^{n+1/2}} \left(\frac{\partial u_h}{\partial t} + \nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\end{aligned} \tag{127}$$

Equivalently, denoting by $\beta_i^{P_E^{n+1/2}}$ the distribution coefficient of node i:

$$\beta_i^{P_E^{n+1/2}} = \frac{\Phi_i^{P_E^{n+1/2}}}{\Phi^{P_E^{n+1/2}}} \tag{128}$$

one must have by construction

$$\sum_{j\in E} \beta_j^{P_E^{n+1/2}} = 1 \tag{129}$$

3. $\forall i \in \mathcal{T}_h$ *assemble the elemental contributions of all $P_E^{n+1/2} \in \mathcal{D}_i$ and compute the nodal values of u^{n+1} by solving the algebraic system*

$$\sum_{P_E^{n+1/2}\in\mathcal{D}_i} \Phi_i^{P_E^{n+1/2}} = 0, \quad \forall i \in \mathcal{T}_h \tag{130}$$

Before proceeding with the analysis of this new prototype, let us give a few remarks. First of all, the above definition introduces a continuous approximation of the unknown (and of flux and source term) in *space and time*. As announced, this section only considers second-order discretizations, in which case it is assumed that u_h has the following particular form:

$$u_h = \frac{t - t^n}{\Delta t} u^{n+1} + \frac{t^{n+1} - t}{\Delta t} u^n \tag{131}$$

Note that in the slab $\Omega \times [t^n, t^{n+1}]$, u_h can be recast as a continuous space-time bilinear interpolant of the data $\{u_i^k\}_{i\in\mathcal{T}_h}^{k=n,n+1}$, with basis

$$L_i^n = \frac{t^{n+1}-t}{\Delta t}\psi_i\,, \quad L_i^{n+1} = \frac{t-t^n}{\Delta t}\psi_i \tag{132}$$

Concerning $\mathcal{F}_h$ and $\mathcal{S}_h$, we will shortly see that, as in the steady case, their choice can be made on the basis of accuracy considerations.

As a second remark, note that the algorithm defined by steps 1–3 constitutes a time-marching procedure, the nodal values at time t^n being considered as given data. A more subtle way to see this is that on a space-time prism no residual is distributed to nodes at the *past* level. We shall discuss this issue in some more detail when introducing the concept of space-time upwinding. Until then, we shall simplify the notation by dropping the superscript $n+1/2$, referring to the prism $E \times [t^n, t^{n+1}]$ simply as P_E. Hence, element residual, local nodal residual, and distribution coefficients will be denoted by Φ^{P_E}, $\Phi_i^{P_E}$, and $\beta_i^{P_E}$ respectively.

Finally, we underline that more general variants of the class of schemes of Definition 10 exist. Perhaps the most important aspect to highlight is that several different ways of approximating the time derivative can be thought of. Since such a general characterization is out of the scope of this paper, we limit ourselves to the comment that, other than approximations obtained via continuous discrete functional representations in space-time, one can resort to the application of a finite difference formula in time to obtain a semidiscrete equivalent of the continuous problem, to be fed into (126) for the computation of the residual. This allows somehow to decouple the approximation of the temporal derivative from the spatial ones (see e.g. the representation used in Ricchiuto, Abgrall and Deconinck, 2007). In addition to this, the integral in time in (126) is not strictly necessary. In particular, one can think of schemes obtained by first discretizing the time derivative, and then distributing a residual defined as the spatial integral of the semidiscrete operator obtained in this way. The first second (and higher) order $\mathcal{RD}$ schemes were actually constructed in this way (Ferrante and Deconinck, 1997; Maerz and Degrez, 1996; Caraeni, 2000). More recent examples can be found in Abgrall, Andrianov and Mezine (2005), Caraeni and Fuchs (2005), De Palma *et al.* (2005), and Rossiello *et al.* (2007). The approach considered here is instead more closely related to the truly space-time formulation of Abgrall and Mezine (2003a), Csík and Deconinck (2002), and Csík, Ricchiuto and Deconinck (2003a).

In the following sections, we review some of the basic properties of the schemes characterized by Definition 10 and give some examples.

5.2.1 Accuracy

This section considers the characterization of the accuracy of the prototype of Definition 10. In the analysis which follows, we make explicit use of the regularity hypothesis on time step and mesh size (second in (8)), so that we have $\Delta t = \mathcal{O}(h)$ and vice versa. The following can be shown (the reader is referred to Rossiello *et al.* (2007) and Ricchiuto, Abgrall and Deconinck (2007), and to Caraeni and Fuchs (2005), Abgrall and Mezine (2003a), and De Palma *et al.* (2005) and references therein).

Proposition 13 (Space-time $\mathcal{RD}$: second order of accuracy) *Given any smooth function $\varphi \in C^1(\Omega \times [0,\, t_f])$, with $\varphi(\cdot, t)$ having compact support on Ω. Given a discretization of the spatial and temporal domain satisfying (8). Given u_h, $\mathcal{F}_h$, and $\mathcal{S}_h$, continuous, second-order accurate space-time interpolants of a smooth exact solution to (4), and of the corresponding exact flux $\mathcal{F}(u)$ and source term $\mathcal{S}(u, x, y, t)$. Then, a space-time $\mathcal{RD}$ verifies the truncation error estimate*

$$TE(u_h, t_f) := \sum_{n=0}^{N}\sum_{i\in\mathcal{T}_h}\varphi_i^{n+1}\sum_{P_E\in\mathcal{D}_i}\Phi_i^{P_E}(u_h) = \mathcal{O}(h^2) \tag{133}$$

provided that the following condition is met

$$\Phi_i^{P_E} = \mathcal{O}(h^4) \tag{134}$$

Note that, even though apparently different, the last condition is consistent with the condition for second order of accuracy at steady state ($\phi_i^E(u_h) = \mathcal{O}(h^3)$ (cf. equation (25)). The extra order of magnitude in (134) is due to the extra integral in time used for the definition of the element residuals, which brings an extra $\mathcal{O}(\Delta t) = \mathcal{O}(h)$ into the analysis (see Rossiello *et al.*, 2007; Ricchiuto, Abgrall and Deconinck, 2007 for details).

Proposition 4.2 gives a criterion to choose the discrete approximations (polynomial interpolant) of the flux and of the source term. Indeed, condition (134) is valid *provided that $\mathcal{F}_h$ and $\mathcal{S}_h$ are second-order accurate*. For a given smooth solution u, an obvious choice is to take $\mathcal{F}_h = \mathcal{F}(u_h)$ and $\mathcal{S}_h = \mathcal{S}(u_h, x, y, t)$. However, we note that the bilinear polynomials

$$\mathcal{F}_h = \frac{t-t^n}{\Delta t}\mathcal{F}^{n+1} + \frac{t^{n+1}-t}{\Delta t}\mathcal{F}^n$$
$$\text{and} \quad \mathcal{S}_h = \frac{t-t^n}{\Delta t}\mathcal{S}^{n+1} + \frac{t^{n+1}-t}{\Delta t}\mathcal{S}^n$$

with $\mathcal{F}^{n+1}$, $\mathcal{F}^n$, $\mathcal{S}^{n+1}$, and $\mathcal{S}^n$ linear in space (cf. equations (131), (9), and (11)), also satisfy this requirement. In

this particular case, the element residual can be explicitly computed as follows:

$$\Phi^{P_E} = \sum_{j\in E} \frac{|E|}{3}(u_j^{n+1} - u_j^n) + \frac{\Delta t}{2}\sum_{j\in E} \frac{\mathcal{F}_j^{n+1} + \mathcal{F}_j^n}{2}\cdot \vec{n}_j - \frac{\Delta t}{2}\sum_{j\in E}\frac{|E|}{3}(\mathcal{S}_j^{n+1} + \mathcal{S}_j^n) \tag{135}$$

In particular, simple arguments can be used to show that, given a smooth exact solution u, for any second-order accurate variable, flux, and source term approximations in space and time, one has (Rossiello *et al.*, 2007; Ricchiuto, Abgrall and Deconinck, 2007)

$$\Phi^{P_E}(u_h) = \mathcal{O}(h^4)$$

Hence, as in the steady case, the following characterization is possible.

Definition 11 ($\mathcal{LP}$ space-time $\mathcal{RD}$ schemes) *A space-time $\mathcal{RD}$ scheme for which $\Phi_i^{P_E} = \beta_i^{P_E}\Phi^{P_E}$, with $\beta_i^{P_E}$ uniformly bounded, that is*

$$\max_{E\in\mathcal{T}_h}\max_{j\in E} \|\beta_j^{P_E}\| < C < \infty \quad \forall\, \Phi^{P_E}, u_h, u_h^0, h, \delta t^n, \dots$$

is said to be ($\mathcal{LP}$). For any given second-order approximation of the variable, the flux, and the source term, a $\mathcal{LP}$ scheme verifies by construction the truncation error estimate (133).

Space-time $\mathcal{RD}$ schemes can also be abstractly represented by introducing the following discrete prototype:

$$\sum_{E\in\mathcal{D}_i}\left(\sum_{j\in E} m_{ij}^E(u_j^{n+1} - u_j^n) + \phi_i\right) = 0 \quad \forall\, i \in \mathcal{T}_h \tag{136}$$

where ϕ_i represents any splitting of the spatial part of the residual

$$\sum_{j\in E}\phi_j = \int_{t^n}^{t^{n+1}}\int_E \left(\nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

and with m_{ij}^E a mass matrix respecting the consistency constraints:

$$\sum_{i\in E}\beta_i^M = 1 \quad \text{with} \quad \beta_i^M = \frac{1}{|E|}\sum_{j\in E} m_{ij}^E \tag{137}$$

where the superscript M stands for mass matrix. This representation shows another feature in common with finite-element methods: a consistent discretization in space, naturally leads to the appearance of a mass matrix multiplying the time derivative. The first examples of second- and third-order schemes of this type are due to the independent work of Caraeni (2000), Caraeni and Fuchs (2002), Caraeni and Fuchs (2005), Maerz and Degrez (1996), and Ferrante and Deconinck (1997).

Concerning the form of the mass matrix, a very interesting analysis, based on geometrical arguments, can be found in De Palma *et al.* (2005). For $\mathcal{LP}$ schemes, one can show the following particularly simple form:

$$m_{ij}^E = \frac{|E|}{3}\begin{bmatrix} \beta_1^{P_E} & \beta_1^{P_E} & \beta_1^{P_E} \\ \beta_2^{P_E} & \beta_2^{P_E} & \beta_2^{P_E} \\ \beta_3^{P_E} & \beta_3^{P_E} & \beta_3^{P_E} \end{bmatrix}$$

having denoted the nodes of E by $\{1, 2, 3\}$. In particular, in this case $\beta_i^M = \beta_i^{P_E}$. Other examples of mass matrices are given hereafter.

5.2.2 *Examples of mass matrices: finite-element schemes*

A well-known member of the class of schemes defined by (136) is the Galerkin $\mathcal{FE}$ scheme. In the space-time slab $\Omega\times[t^n, t^{n+1}]$, it is defined by

$$\int_{t^n}^{t^{n+1}}\int_\Omega \psi_i\frac{\partial u_h}{\partial t}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{t^n}^{t^{n+1}}\int_\Omega \psi_i \times\left(\nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = 0, \quad \forall i\in\mathcal{T}_h \tag{138}$$

If ψ_i denotes the continuous piecewise linear shape function, we end up with a scheme formally identical to (136) with the Galerkin mass matrix given by

$$m_{ij}^E = m_{ij}^{\mathrm{G}} = \frac{|E|}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Note that, strictly speaking, this is not a truly space-time Galerkin scheme, the test function being the standard linear shape function in space, and not the space-time bilinear polynomials (132), used to contruct (131).

The streamline dissipation Galerkin scheme with stabilization parameter τ can be derived in a similar fashion:

$$\int_{t^n}^{t^{n+1}}\int_\Omega \psi_i\frac{\partial u_h}{\partial t}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \int_{t^n}^{t^{n+1}}\int_\Omega \psi_i\left(\nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \sum_{E\in\mathcal{T}_h}\int_{t^n}^{t^{n+1}}\int_E \tau\,\widetilde{a}\cdot\nabla\psi_i\frac{\partial u_h}{\partial t}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \sum_{E\in\mathcal{T}_h}\int_{t^n}^{t^{n+1}}\int_E \tau\,\widetilde{a}\cdot\nabla\psi_i\left(\nabla\cdot\mathcal{F}_h - \mathcal{S}_h\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = 0 \tag{139}$$

with $\widetilde{a}$ being a properly chosen average of the flux Jacobian (cf. equation (115)). As before, we obtain a scheme formally identical to (136) with

$$m_{ij}^{\text{SD-G}} = \frac{1}{12}\begin{bmatrix} 2|E|+2\tau k_1 & |E|+2\tau k_1 & |E|+2\tau k_1 \\ |E|+2\tau k_2 & 2|E|+2\tau k_2 & |E|+2\tau k_2 \\ |E|+2\tau k_3 & |E|+2\tau k_3 & 2|E|+2\tau k_3 \end{bmatrix}$$

We observe that both for the Galerkin scheme and for the SD-G scheme, formulation (31) is obtained by substituting to the mass matrix of the schemes, the *lumped* mass matrix obtained as

$$m_{ij}^{\text{lumped}} = \delta_{ij} \sum_{k\in E} m_{ik}^{E} = \delta_{ij}\frac{|E|}{3}$$

The mass lumping procedure introduces an inconsistency, ultimately spoiling the spatial accuracy of the schemes.

5.2.3 *Examples of mass matrices: a $\mathcal{RD}$ Taylor–Galerkin procedure and the second-order LW scheme*

For the case of the advection equation (5) with zero source term, we show the construction of a consistent second-order cell-vertex $\mathcal{RD}$ LW scheme. We start with the Taylor expansion in time (Paillère, 1995; Roe, 1987):

$$u^{n+1} = u^n + \left(\frac{\partial u}{\partial t}\right)^n \Delta t + \frac{\Delta t^2}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)^n + \mathcal{O}(\Delta t^3)$$

For linear homogeneous scalar advection, one has

$$\frac{\partial u}{\partial t} = -\nabla\cdot(\vec{a}\,u) \quad \text{and} \quad \frac{\partial^2 u}{\partial t^2} = \nabla\cdot\big(\vec{a}\nabla\cdot(\vec{a}\,u)\big)$$

hence

$$\frac{u^{n+1}-u^n}{\Delta t} + \nabla\cdot(\vec{a}\,u)^n - \frac{\Delta t}{2}\nabla\cdot\big(\vec{a}\nabla\cdot(\vec{a}\,u)\big)^n = \mathcal{O}(\Delta t^2)$$

which is a semidiscrete second-order accurate equivalent of the time-dependent advection equation. Neglecting terms of $\mathcal{O} \geq \Delta t^2$ and discretizing the resulting expression with Galerkin $\mathcal{FE}$ leads to the well-known Taylor–Galerkin scheme (Donea and Huerta, 2003). The $\mathcal{RD}$ analog is usually obtained by integrating the last expression over the median dual cell S_i:

$$\int_{S_i}\frac{u^{n+1}-u^n}{\Delta t}\,\mathrm{d}x\,\mathrm{d}y + \int_{S_i}\nabla\cdot(\vec{a}\,u)^n\,\mathrm{d}x\,\mathrm{d}y - \frac{\Delta t}{2}\int_{S_i}\nabla\cdot\big(\vec{a}\nabla\cdot(\vec{a}\,u)\big)^n\,\mathrm{d}x\,\mathrm{d}y = 0$$

which we recast as

$$\sum_{E\in\mathcal{D}_i}\left(\int_{S_i\cap E}\frac{u^{n+1}-u^n}{\Delta t}\,\mathrm{d}x\,\mathrm{d}y + \int_{S_i\cap E}\nabla\cdot(\vec{a}\,u)^n\,\mathrm{d}x\,\mathrm{d}y - \frac{\Delta t}{2}\oint_{\partial S_i\cap E}\nabla\cdot(\vec{a}\,u)^n\vec{a}\cdot\hat{n}\,\mathrm{d}l\right) = 0 \tag{140}$$

One easily checks that for u_h given by (9), and due to the definition of S_i (see also Figure 10),

$$\int_{S_i\cap E}\nabla\cdot(\vec{a}\,u)^n\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{3}\phi^E(u^n)$$
$$\frac{\Delta t}{2}\oint_{\partial S_i\cap E}\nabla\cdot(\vec{a}\,u)^n\vec{a}\cdot\hat{n}\,\mathrm{d}l = -\frac{\Delta t\,k_i}{2|E|}\phi^E(u^n)$$

with $\phi^E(u^n)$ as in (52) (with $\mathcal{S}=0$). Integrating the first term in (140) exactly with respect to a piecewise linear approximation of u^{n+1} and u^n of the type (9), we arrive at the LW scheme

$$\sum_{E\in\mathcal{D}_i}\left(\sum_{j\in E} m_{ij}^{\text{LW}}(u_j^{n+1}-u_j^n) + \Delta t\beta_i^{\text{LW}}\phi^E(u^n)\right) = 0, \quad \forall i\in\mathcal{T}_h$$
$$\beta_i^{\text{LW}} = \frac{1}{3} + \frac{\Delta t\,k_i}{2|E|} \tag{141}$$

where the *consistent $\mathcal{RD}$ LW mass matrix* m_{ij}^{LW} is given by

$$m_{ij}^{\text{LW}} = \frac{|E|}{108}\begin{bmatrix} 22 & 7 & 7 \\ 7 & 22 & 7 \\ 7 & 7 & 22 \end{bmatrix} \tag{142}$$

Scheme (141) is fully consistent with a second-order approximation of the solution in space and, not surprisingly, features a nondiagonal mass matrix. The LW scheme traditionally encountered in literature (Paillère, 1995; Roe, 1987;

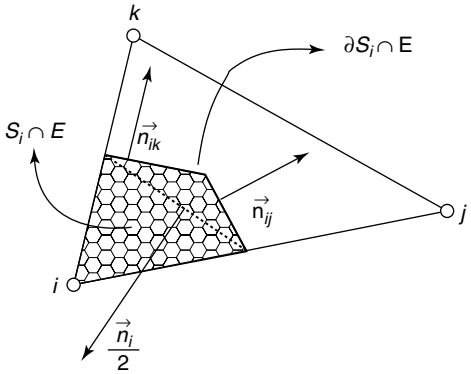

Figure 10. LW scheme: geometry of the construction.

Hubbard and Roe, 2000; De Palma, Pascazio and Napolitano, 2001) is obtained from this consistent discretization after lumping of m_{ij}^{LW}, yielding

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|S_i|} \sum_{E \in \mathcal{D}_i} \beta_i^{\mathrm{LW}} \phi^E(u^n), \quad \forall i \in \mathcal{T}_h \qquad (143)$$

As in the case of the Galerkin and SD-G schemes, one would expect the lumping of the mass matrix to lead to a first-order discretization. Surprisingly, this inconsistency has never been observed in practice, scheme (143) having had some success in literature (Hubbard and Roe, 2000; De Palma, Pascazio and Napolitano, 2001; De Palma *et al.*, 2005). With the exception of the recent results reported in De Palma *et al.* (2005), we remark however that most of the numerical tests presented in the references were performed on regular grids, on which error cancellation might occur. For triangular grids obtained by cutting a Cartesian grid with uniformly right-running diagonals, this has been shown in Ricchiuto and Deconinck (1999), where the modified equation of scheme (143) has been derived, showing second order of accuracy.

The results presented in De Palma *et al.* (2005) on unstructured grids seem to fall out of the last remarks. An explanation of these results might be obtained by adapting to the time-dependent case the observations done for the treatment of source terms in Section 5.2.1. However, at the moment, no formal evidence has been given to show that the *inconsistent* LW scheme (143) is second-order accurate.

5.2.4 *Monotonicity*

The schemes of Definition 10 are inherently implicit. Their monotonicity will generally depend on the form of the mass matrix. Generally speaking, the idea is that if the spatial part of (136) defines a LED scheme and if the mass matrix is an $\mathcal{M}$-matrix (Berman and Plemmons, 1979), then, upon its inversion, one would end up with a scheme that is still LED, hence respecting a discrete maximum principle.

For the particular case of Definition 10, with u_h taken as in (131), one can characterize this property making use of the analysis made in Section 3.3. Let us then consider the homogeneous advection equation obtained by (5) with $\mathcal{S} = 0$. When u_h is taken as in (131), the residual on a space-time prism P_E can be written as (cf. equation (135))

$$\begin{aligned}\Phi^{P_E} &= \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u_h}{\partial t} + \vec{a} \cdot \nabla u_h \right) \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t \\ &= \sum_{j \in E} \frac{|E|}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{j \in E} (k_j u_j^n + k_j u_j^{n+1})\end{aligned}$$

The last expression can equivalently be recast as

$$\Phi^{P_E} = \sum_{j \in E} \frac{|E|}{3} (u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} (\phi^E(u^n) + \phi^E(u^{n+1}))$$

where $\phi^E(u^n)$ and $\phi^E(u^{n+1})$ are the steady scalar element residuals of equation (52) (with $\mathcal{S} = 0$), evaluated at time t^n and t^{n+1} respectively. Suppose now that a LED splitting of the steady element residual is given and denote the corresponding local nodal residuals by $\{\phi_j^{\mathrm{LED}}\}_{j \in E}$. A trivial splitting of Φ^{P_E} can then be obtained as

$$\Phi_i^{\mathrm{LED}} = \frac{|E|}{3} (u_i^{n+1} - u_i^n) + \frac{\Delta t}{2} (\phi_i^{\mathrm{LED}}(u^n) + \phi_i^{\mathrm{LED}}(u^{n+1}))$$

One immediately sees from the last expression that this approach is equivalent to the one of Definition 1, when the discretization in time is performed with the $\mathcal{CN}$ scheme, or equivalently, in the case of the advection equation, with the trapezium scheme. This remark, combined with proposition 4, leads to the following result.

Proposition 14 (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* (16) *with the trapezium or $\mathcal{CN}$ scheme. The positivity of the resulting discretization is constrained by the time-step restrictions of proposition 4.*

The last proposition defines a particular class of monotone schemes. However, we have still not exploited to its maximum the space-time nature of the discretization defined by Definition 10. This will be done in the following section.

5.3 Multidimensional upwinding in space-time

The objective of this section is to try to make explicit use of the local space-time geometry of the prism $E \times [t^n, t^{n+1}]$ to construct schemes that incorporate, at the discrete level, the directional propagation of the information that is typical of solutions to (5). The idea is to rewrite the element residual as

$$\begin{aligned}\Phi^{P_E} &= \sum_{j \in E} \left(\frac{\Delta t k_j}{2} + \frac{|E|}{3} \right) u_j^{n+1} + \sum_{j \in E} \left(\frac{\Delta t k_j}{2} - \frac{|E|}{3} \right) u_j^n \\ &= \sum_{j \in E} \overline{k}_j u_j^{n+1} + \sum_{j \in E} \widehat{k}_j u_j^n \qquad (144)\end{aligned}$$

Introducing the *space-time flux* $(\vec{a}u, u) \in \mathbb{R}^2 \times \mathbb{R}$, we can show that the $\overline{k}_j$ and $\widehat{k}_j$ parameters, implicitly defined

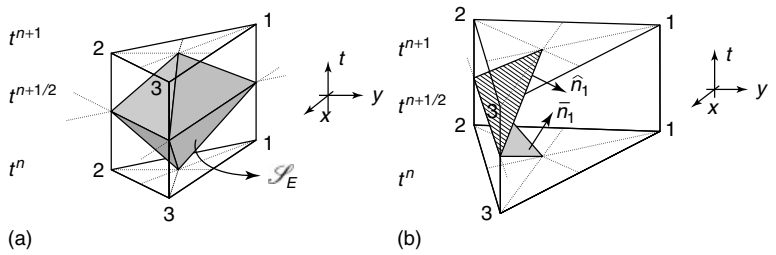

Figure 11. Closed shell in $E \times [t^n, t^{n+1}]$ (a), and space-time directions $\overline{n}_1$ and $\widehat{n}_1$ (b).

by (144), 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 $E \times [t^n, t^{n+1}]$. To do this, we consider the shell $\mathscr{S}_E$ formed by joining the gravity centers of E at times t^n and t^{n+1} with the nodes of the element at time $t^{n+1/2} = t^n + (t^{n+1} - t^n)/2$ (Figure 11a). We can associate to each node of the prism the face of $\mathscr{S}_E$ opposite to it, as illustrated in Figure 11(b) for node 1. With reference to this last picture, we introduce the space-time vectors $\overline{n}_1$ and $\widehat{n}_1$, normal to the faces of $\mathscr{S}_E$ opposite to node 1, pointing inward with respect to the shell, and scaled by the area of the faces.

Simple geometry shows that

$$\overline{k}_1 = \overline{n}_1 \cdot (\vec{a}, 1) \quad \text{and} \quad \widehat{k}_1 = \widehat{n}_1 \cdot (\vec{a}, 1)$$

Since $(\vec{a}, 1)$ is the direction of a characteristic line cutting through the prism, we deduce that $\overline{k}_1$ and $\widehat{k}_1$ are the projections of the direction of the characteristic onto $\overline{n}_1$ and $\widehat{n}_1$. For the exact solution of the advection equation, all the information propagates along $(\vec{a}, 1)$. We have 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^{P_E} only if $\overline{k}_1 > 0$. This philosophy is at the basis of the schemes proposed in Csík and Deconinck (2002) and Csík, Ricchiuto and Deconinck (2003a), the case of prismatic space-time elements being discussed in Csík, Ricchiuto and Deconinck (2003a). In particular, one can introduce *space-time inflow and outflow* states defined as

$$\begin{aligned}\overline{u}_{\text{in}} &= \sum_{j\in E}\left(\sum_{j\in E}(\overline{k}_j^- + \widehat{k}_j^-)\right)^{-1}(\overline{k}_j^- u_j^{n+1} + \widehat{k}_j^- u_j^n) \\ &= -\sum_{j\in E}\overline{N}(\overline{k}_j^- u_j^{n+1} + \widehat{k}_j^- u_j^n) \end{aligned} \tag{145}$$

and

$$\begin{aligned}\overline{u}_{\text{out}} &= \sum_{j\in E}\left(\sum_{j\in E}(\overline{k}_j^+ + \widehat{k}_j^+)\right)^{-1}(\overline{k}_j^+ u_j^{n+1} + \widehat{k}_j^+ u_j^n) \\ &= \sum_{j\in E}\overline{N}(\overline{k}_j^+ u_j^{n+1} + \widehat{k}_j^+ u_j^n) \end{aligned} \tag{146}$$

with

$$\overline{N} = \left(\sum_{j\in E}(\overline{k}_j^+ + \widehat{k}_j^+)\right)^{-1} \tag{147}$$

This notation allows to express the residual as

$$\Phi^{P_E} = \left(\sum_{j\in E}(\overline{k}_j^+ + \widehat{k}_j^+)\right)(\overline{u}_{\text{out}} - \overline{u}_{\text{in}}) \tag{148}$$

The last equations show the analogy with a one-dimensional balance along the characteristic line ζ intersecting the prism $E \times [t^n, t^{n+1}]$ in $\overline{u}_{\text{out}}$ and $\overline{u}_{\text{in}}$. Note however that since the $\widehat{k}_j$ are not necessarily all negative, $\overline{u}_{\text{in}}$ does not necessarily lie on the plane $t = t^n$. Similarly, $\overline{u}_{\text{out}}$ does not necessarily lie on the plane $t = t^{n+1}$. In general, one will have a configuration as, for example, the one in Figure 12 (Even though in the most general situation both $\overline{u}_{\text{out}}$ and $\overline{u}_{\text{in}}$ are inside the prism).

These concepts allow a natural extension of the $\mathcal{MU}$ idea to the space-time framework. To be able to do this, first we have to enlarge the class of schemes we consider. As already remarked, Definition 10 gives a time-marching procedure allowing to compute the unknown at time t^{n+1}, given its nodal values at time t^n. Here, we *suppose instead to be solving on the entire space-time domain at once*, on a discretization that 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

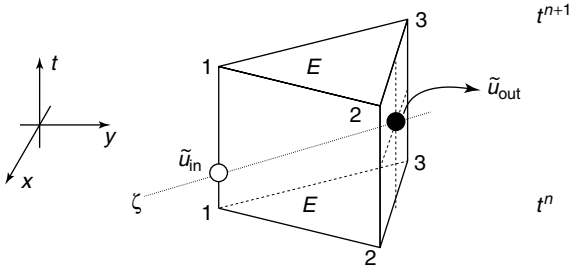

Figure 12. Space-time inflow and outflow states.

this case, the fully discrete analog of (5) can be written as

$$\sum_{E\in\mathcal{D}_i} \Phi_{i,n}^{P_E^{n-1/2}} + \sum_{E\in\mathcal{D}_i} \Phi_{i,n}^{P_E^{n+1/2}} = 0, \quad \forall i \in \mathcal{T}_h, \forall n = 2, M-1$$

$$\sum_{E\in\mathcal{D}_i} \Phi_{i,M}^{P_E^{M-1/2}} = 0, \quad \forall i \in \mathcal{T}_h$$

where $\forall E \in \mathcal{T}_h$ and $\forall n = 1, M-1$

$$\sum_{j\in E} \left(\Phi_{j,n-1}^{P_E^{n-1/2}} + \Phi_{j,n}^{P_E^{n-1/2}}\right) = \Phi^{P_E^{n-1/2}}$$

$$\sum_{j\in E} \left(\Phi_{j,n}^{P_E^{n+1/2}} + \Phi_{j,n+1}^{P_E^{n+1/2}}\right) = \Phi^{P_E^{n+1/2}}$$

where

$$\Phi^{P_E^{n-1/2}} = \int_{t^{n-1}}^{t^n} \int_E \left(\frac{\partial u_h}{\partial t} + \vec{a}\cdot\nabla u_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

$$\text{and} \quad \Phi^{P_E^{n+1/2}} = \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial u_h}{\partial t} + \vec{a}\cdot\nabla u_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

So $\Phi_{i,n}^{P_E^{n+1/2}}$ represents the fraction of $\Phi^{P_E^{n+1/2}}$ distributed to the node i lying in the time plane $t = t^n$, as illustrated on Figure 13. We give the following definition of a space-time multidimensional upwind (ST-$\mathcal{MU}$) scheme.

Definition 12 (ST-$\mathcal{MU}$ scheme) *A space-time $\mathcal{RD}$ scheme is ST-$\mathcal{MU}$ if in the prism $P_E^{n+1/2} = E \times [t^n, t^{n+1}]$*

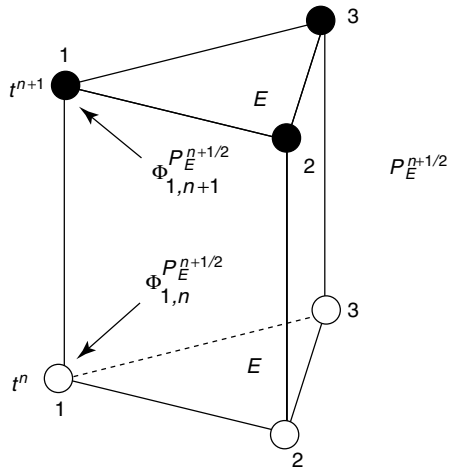

Figure 13. Local nodal residuals $\Phi_{1,n+1}^{P_E^{n+1/2}}$ and $\Phi_{1,n}^{P_E^{n+1/2}}$ on the prism $P_E^{n+1/2}$.

$$\overline{k}_j \le 0 \implies \Phi_{j,n+1}^{P_E^{n+1/2}} = 0$$

$$\widehat{k}_j \le 0 \implies \Phi_{j,n}^{P_E^{n+1/2}} = 0$$

Proposition 15 (Space-time-$\mathcal{MU}$ schemes and time marching) *A ST-$\mathcal{MU}$ scheme defines a time-marching procedure if*

$$\Delta t = t^{n+1} - t^n \le \min_{E\in\mathcal{T}_h} \min_{j\in E} \frac{2|E|}{3\,k_j^+}, \quad \forall\, n = 1, M-1 \quad (149)$$

Proof. Owing to (149), $\widehat{k}_j^+ = 0$ in all the elements, and in all space-time slabs. Hence, in every space-time slab $\Omega \times [t^n, t^{n+1}]$, a ST-$\mathcal{MU}$ scheme will not distribute any residual to the nodes at time t^n, decoupling the values of u_h in these nodes from its values at time t^{n+1}, thus yielding a true time-marching procedure. □

In Csík and Deconinck (2002) and Csík, Ricchiuto and Deconinck (2003a), condition (149) is called the *past-shield* condition. 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 (or $\mathcal{CN}$) time integration (see equation (87)). This condition allows to recast space-time-$\mathcal{MU}$ schemes into the framework of Definition 10. In the following we will always assume that (149) is satisfied. This allows to simplify our notation, going back to the labeling $\Phi_i^{P_E}$, and Φ^{P_E} for the local nodal residuals and element residual respectively, so as to have uniform labeling with the previous sections. No confusion is generated, since (149) guarantees that the characterization of Definition 10 is valid, and only the nodal values of u^{n+1} are to be computed in $\Omega \times [t^n, t^{n+1}]$.

Hereafter we give some examples of upwind and space-time upwind schemes.

5.3.1 *Upwind and space-time upwind $\mathcal{RD}$: LDA schemes*

Several extensions of the LDA scheme to the space-time framework exist. One of these resorts to an analogy with finite-element PG schemes, thus introducing a consistent mass matrix (Maerz and Degrez, 1996; Ferrante and Deconinck, 1997; Abgrall and Mezine, 2003a). According to this analogy, a consistent extension of the LDA scheme is obtained as

$$\Phi_i^{\text{LDA-PG}} = \int_{t^n}^{t^{n+1}} \int_E \left(\psi_i + \left(\beta_i^{\text{LDA}} - \frac{1}{3}\right)\right) \times \left(\frac{\partial u_h}{\partial t} + \vec{a}\cdot\nabla u_h\right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

with ψ_i the piecewise linear basis functions (11). By assuming u_h to have a bilinear variation, we get

$$\Phi_i^{\text{LDA-PG}} = \sum_{j\in E}\left(m_{ij}^{\text{LDA-PG}}(u_j^{n+1}-u_j^n) + \frac{\Delta t}{2}\beta_i^{\text{LDA}}k_j(u_j^{n+1}+u_j^n)\right)$$
$$m_{ij}^{\text{LDA-PG}} = \frac{|E|}{3}\left(\beta_i^{\text{LDA}} - \frac{1}{12} + \frac{\delta_{ij}}{4}\right) \tag{150}$$

with δ_{ij} Kroenecker's delta.

A different formulation is obtained by distributing the space-time residual as

$$\Phi_i^{\text{LDA}} = k_i^+ N\Phi^h = \beta_i^{\text{LDA}}\Phi^{P_E} \tag{151}$$

with N as in (77). Scheme (151) is equivalent to the one originally proposed in Caraeni (2000) (see also Caraeni and Fuchs, 2002; Caraeni and Fuchs, 2005).

Finally one can use the space-time $\mathcal{MU}$ variant of the LDA scheme given by

$$\Phi_i^{\text{ST-LDA}} = \overline{k}_i^+\overline{N}\,\Phi^{P_E} = \overline{\beta}_i^{\text{ST-LDA}}\Phi^{P_E} \tag{152}$$

where now, owing to the satisfaction of (149), the parameter $\overline{N}$ is given by

$$\overline{N} = \left(\sum_{j\in E}\overline{k}_j^+\right)^{-1} \tag{153}$$

By construction, the LDA-PG, LDA, and space-time low diffusion A (ST-LDA) schemes all respect the accuracy condition of Proposition 13, and hence they are formally second-order accurate.

5.3.2 *Upwind and space-time upwind $\mathcal{RD}$: N schemes*

Two extensions of the N scheme to the space-time framework exist in literature. The first follows from proposition 14 and is defined by the space-time local nodal residual

$$\Phi_i^{\text{N}} = \frac{|E|}{3}(u_i^{n+1}-u_i^n) + \frac{\Delta t}{2}k_i^+(u_i^n - u_{\text{in}}^n) + \frac{\Delta t}{2}k_i^+(u_i^{n+1}-u_{\text{in}}^{n+1}) \tag{154}$$

This is the positive first-order space-time N (ST-N) scheme as proposed in Abgrall and Mezine (2003a). 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

$$\Phi_i^{\text{ST-N}} = \overline{k}_i^+(u_i^{n+1} - \overline{u}_{\text{in}}) \tag{155}$$

with $\overline{u}_{\text{in}}$ as in (145). The satisfaction of the past-shield condition guarantees that the ST-N scheme (155) satisfies the consistency condition (127). Moreover, it has, as the scheme defined by (86), a subelement LED character, in space-time, which formally ensures the satisfaction of the local space-time discrete maximum principle (40). As the N and the LDA schemes, the ST-N and ST-LDA schemes are linked by

$$\Phi_i^{\text{ST-N}} = \Phi_i^{\text{ST-LDA}} + d_i^{\text{ST-N}} \tag{156}$$

where $d_i^{\text{ST}-\text{N}}$ is a space-time dissipation term given by

$$d_i^{\text{ST-N}} = \sum_{j\in E}\overline{k}_i^+\overline{Nk}_j^+(u_i^{n+1}-u_j^{n+1}) \tag{157}$$

The space-time nature of this term is such that the ST-N schemes is generally extremely more dissipative than scheme (154), as confirmed by the results available in literature (Ricchiuto, 2005; Ricchiuto, Csík and Deconinck, 2005; Csík, Ricchiuto and Deconinck, 2003a; Ricchiuto, Abgrall and Deconinck, 2007). Note also that, while the N scheme of Abgrall and Mezine (2003a) reduces to the standard N scheme at steady state, the same is not true for the ST-N scheme.

5.3.3 *Nonlinear schemes*

The techniques described in Section 4.4 can also be used in time-dependent computations, in the space-time framework. In this case, little is known about the L^2 stability of the resulting schemes (however, see the recent work of Ricchiuto and Abgrall, 2006).

$\mathcal{LP}$ space-time schemes can be obtained by applying mapping (106) either to the N scheme (154) or to the ST-N scheme (155). The LN scheme and the limited space-time N (LST-N) scheme are $\mathcal{LP}$ by construction, and *inherit* positivity from the linear schemes since

$$\Phi_i^{\text{limited}} = \gamma_i\Phi_i^{\text{linear}}, \quad \gamma_i \geq 0$$

Strictly speaking, last relation makes sense only if the local positivity of the linear scheme is ensured.

5.4 Nonlinear conservation laws

The extension to nonlinear conservation laws is achieved exactly as in the steady case. For example, in the case of the space-time schemes on bilinear prismatic elements, one can give the following definition.

Definition 13 (Conservative space-time $\mathcal{RD}$) *A space-time $\mathcal{RD}$ scheme is conservative if there exist a continuous approximation of the unknown u^h, of the flux $\mathcal{F}_h$ and of the source term $\mathcal{S}_h$, such that*

$$\Phi^{P_E} = \int_E (u_h(t^{n+1}) - u_h(t^n))\,\mathrm{d}x\,\mathrm{d}y + \int_{t^n}^{t^{n+1}} \oint_{\partial E} \mathcal{F}_h \cdot \hat{n}\,\mathrm{d}l\,\mathrm{d}t - \int_{t^n}^{t^{n+1}} \int_E \mathcal{S}_h\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \quad (158)$$

In the homogeneous case, for example, one way to compute the space-time residual is

$$\Phi^{P_E} = \int_{P_E} \left(\frac{\partial u_h}{\partial t} + \vec{a}(u_h) \cdot \nabla u_h \right) \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t = \sum_{j \in E} \frac{|E|}{3}(u_j^{n+1} - u_j^n) + \frac{\Delta t}{2} \sum_{j \in E} \widetilde{k}_j^{n+1} u_j^{n+1} + \frac{\Delta t}{2} \sum_{j \in E} \widetilde{k}_j^n u_j^n$$

with $\widetilde{k}_j^{n+1}$ and $\widetilde{k}_j^n$ still defined by (53), except that they are computed using the mean-value Jacobians

$$\widetilde{a}^{n+1} = \frac{2}{|E|\Delta t} \int_{t^n}^{t^{n+1}} \int_E \frac{t - t^n}{\Delta t}\,\vec{a}(u_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

$$\text{and} \quad \widetilde{a}^n = \frac{2}{|E|\Delta t} \int_{t^n}^{t^{n+1}} \int_E \frac{t^{n+1} - t}{\Delta t}\,\vec{a}(u_h)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t$$

The schemes obtained in this way verify Definition 13 for $\mathcal{F}_h = \mathcal{F}(u_h)$. This approach leads to a straightforward extension to the nonlinear case. However, the need for computing conservative mean-value Jacobians with sufficient accuracy leads to a considerable computational cost, which can be large when going to systems.

A simpler approach is to compute the space-time residual directly as (cf. equation (135))

$$\Phi^{P_E} = \sum_{j \in E} \frac{|E|}{3}(u_j^{n+1} - u_j^n) + \Delta t \sum_{l_j=1}^{3} \mathcal{F}^{l_j} \cdot \vec{n}_{l_j} - \Delta t \sum_{j \in E} \frac{|E|}{3} \mathcal{S}_j \quad (159)$$

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

$$\mathcal{F}^{l_j} = \frac{1}{2} \sum_{p=1}^{\mathrm{N_C}} \omega_p \mathcal{F}\big(u^n(x_p, y_p)\big) + \frac{1}{2} \sum_{p=1}^{\mathrm{N_C}} \omega_p \mathcal{F}\big(u^{n+1}(x_p, y_p)\big)(x_p, y_p) \in l_j \quad (160)$$

with $u^n(x, y)$ and $u^{n+1}(x, y)$ piecewise linear. The condition for second order of accuracy $\Phi^{P_E} = \mathcal{O}(h^4)$ is already fulfilled by an exact integration assuming a piecewise linear variation of the flux, leading to (135). Conservative $\mathcal{LP}$ schemes are second-order accurate in space and time. Clearly (159, 160) alone do not give a nodal approximation of (4). A distribution strategy has to be formulated. This is easily achieved by combining (159, 160) with the definitions of the LDA and N schemes given in Section 4.2.3 (see also Section 4.3.3) and with the conservative formulation based on contour integration of Section 4.5.1 (see Ricchiuto, Csík and Deconinck, 2005; Ricchiuto, Abgrall and Deconinck, 2007 for details).

6 EXTENSION TO SYSTEMS AND APPLICATIONS

We briefly sketch the extension of $\mathcal{RD}$ schemes to the discretization of systems of conservation laws and show numerical results, representative of some of the latest developments on $\mathcal{RD}$.

In the first paragraphs, we describe the *matrix* variant of $\mathcal{RD}$ schemes, which is by far the most used in published literature. Other approaches exist, for which we refer the reader to the references given in the introduction.

The second part of this section shows the potential of the schemes in computing in an accurate and nonoscillatory way complex solutions to conservation laws.

6.1 Matrix residual distribution schemes

Two approaches exist to apply $\mathcal{RD}$ schemes for the discretization of hyperbolic systems of conservation laws. One way to achieve this extension is to use the mathematical structure of the system, trying to split the coupled equations into a subset of more or less uncoupled PDEs. For the Euler equations of gas dynamics, the most successful version of this approach allows, in the steady two-dimensional case, to split the system in four uncoupled *scalar* transport equations (total enthalpy, entropy, and two Riemann invariants) in the supersonic case, while in the subsonic case one can split the systems in two *scalar* transport equations (total enthalpy, entropy) plus a coupled elliptic subsystem (hyperbolic elliptic splitting) (Paillère, 1995; Nishikawa, Rad and Roe, 2001). The advantage of this technique is that it allows to discretize each split equation (or system) with a different scheme. In the hyperbolic elliptic splitting of Nishikawa, Rad and Roe (2001), for example, the scalar equations are solved by a high-order monotone $\mathcal{RD}$ scheme, while the elliptic subsystem is discretized with a

least squares approach or with a LW (matrix) distribution scheme. This approach has been widely used in the early years of the development of $\mathcal{RD}$ schemes for compressible flow simulations (Struijs, Deconinck and Roe, 1991; Paillère, 1995; Mesaros, 1995).

Although quite powerful, this technique has some important limits in the fact that it does not generalize to the three-dimensional (and time-dependent) case and that it is tailored to a particular set of equations.

A more general approach is the so-called *matrix* approach, initially proposed in van der Weide and Deconinck (1996) and van der Weide *et al.* (1999). The idea underlying matrix $\mathcal{RD}$ schemes is to extend *formally* scalar distribution schemes to systems, by replacing vector Jacobians by matrices, whose eigenstructure is used to obtain an upwind discretization. The extension of this approach to three space dimensions, to the time-dependent case, and to any system of equations does not present any difficulty. In the next paragraphs, we briefly review the matrix $\mathcal{RD}$ approach. The reader is referred to the bibliography for a more extensive overview on the subject.

6.1.1 Linear hyperbolic systems of PDEs

Consider the linear symmetric hyperbolic system of PDEs

$$\begin{aligned} &\frac{\partial \mathbf{u}}{\partial t} + A_1 \frac{\partial \mathbf{u}}{\partial x} + A_2 \frac{\partial \mathbf{u}}{\partial y} = 0 \\ \text{on} \quad &\Omega_T = \Omega \times [0, t_f] \subset \mathbb{R}^d \times \mathbb{R}^+ \end{aligned} \tag{161}$$

with $\mathbf{u}$ being a vector of m unknowns. To illustrate the basic idea of the matrix approach, it suffices to assume on $\mathcal{T}_h$, an unstructured triangulation of Ω, a piecewise linear discrete representation of the unknown $\mathbf{u}_h$ of type (9), and compute on $E \in \mathcal{T}_h$ the spatial residual

$$\boldsymbol{\phi}^E = \int_E \left(A_1 \frac{\partial \mathbf{u}_h}{\partial x} + A_2 \frac{\partial \mathbf{u}_h}{\partial y} \right) \mathrm{d}x \, \mathrm{d}y$$

Straightforward calculations immediately lead to

$$\boldsymbol{\phi}^E = \sum_{j \in E} K_j \mathbf{u}_j, \quad K_j = \frac{1}{2} \left(A_1 n_{jx} + A_2 n_{jy} \right) \tag{162}$$

which is formally identical to (52) in the homogeneous case, except that in (162) the k_j parameters have been replaced by matrices. In particular, the system being hyperbolic, the K_j matrices admit a decomposition in a positive and a negative part, in the standard matrix sense

$$K_j^{\pm} = \frac{1}{2} \left(K_j \pm |K_j| \right), \quad |K_j| = R_j |\Lambda_j| R_j^{-1} \tag{163}$$

with R_j the matrix of the right eigenvectors of K_j, Λ_j the diagonal matrix of the eigenvalues, and $|\Lambda_j|$ the diagonal matrix of the absolute values of the eigenvalues of K_j. With this basic notation, one can easily write the matrix variant of the distribution schemes described in the previous sections.

In particular, matrix extensions of the LDA and of the N schemes are defined by the split residuals

$$\boldsymbol{\phi}_i^{\mathrm{LDA}} = \boldsymbol{\beta}_i^{\mathrm{LDA}} \boldsymbol{\phi}^E, \quad \boldsymbol{\beta}_i^{\mathrm{LDA}} = K_i^+ N, \quad N = \left(\sum_{j \in E} K_j^+ \right)^{-1} \tag{164}$$

and

$$\boldsymbol{\phi}_i^{\mathrm{N}} = K_i^+ (\mathbf{u}_i - \mathbf{u}_{\mathrm{in}}), \quad \mathbf{u}_{\mathrm{in}} = -N \sum_{j \in E} K_j^- \mathbf{u}_j \tag{165}$$

For symmetrizable systems, the existence of matrix products of the type $K_i^+ N$ is proved in Abgrall (2001) and Abgrall and Mezine (2003b).

The extension to the time-dependent case is obtained in a similar fashion, the space-time schemes of Definition 10 admitting a natural matrix formulation. For example, on bilinear elements the space-time residual is easily shown to be

$$\begin{aligned} \boldsymbol{\Phi}^{P_E} &= \int_{t^n}^{t^{n+1}} \int_E \left(\frac{\partial \mathbf{u}_h}{\partial t} + A_1 \frac{\partial \mathbf{u}_h}{\partial x} + A_2 \frac{\partial \mathbf{u}_h}{\partial y} \right) \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t \\ &= \sum_{j \in E} \frac{|E|}{3} \left(\mathbf{u}_j^{n+1} - \mathbf{u}_j^n \right) + \frac{\Delta t}{2} \sum_{j \in E} \left(K_j \mathbf{u}_j^{n+1} + K_j \mathbf{u}_j^n \right) \end{aligned} \tag{166}$$

Nodal discrete equations are obtained by splitting in every $P_E \in \mathcal{T}_h$ the residual (166), as described by Definition 10. For example, with the obvious meaning of the symbols, matrix variants of the N scheme (154) and of the ST-N scheme (155) are defined by

$$\begin{aligned} \boldsymbol{\Phi}_i^{\mathrm{N}} = \frac{|E|}{3} (\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \frac{\Delta t}{2} K_i^+ (\mathbf{u}_i^n - \mathbf{u}_{\mathrm{in}}^n) \\ + \frac{\Delta t}{2} K_i^+ (\mathbf{u}_i^{n+1} - \mathbf{u}_{\mathrm{in}}^{n+1}) \end{aligned} \tag{167}$$

and by

$$\begin{aligned} \boldsymbol{\Phi}_i^{\text{ST-N}} &= \widetilde{K}_i^+ (\mathbf{u}_i^{n+1} - \widetilde{\mathbf{u}}_{\mathrm{in}}) \\ \widetilde{\mathbf{u}}_{\mathrm{in}} &= -\widetilde{N} \sum_{j \in E} (\widetilde{K}_j^- \mathbf{u}_j^{n+1} + \widehat{K}_j^- \mathbf{u}_j^n) \end{aligned} \tag{168}$$

where

$$\widetilde{N} = \left(\sum_{j \in E} (\widetilde{K}_j^+ + \widehat{K}_j^+) \right)^{-1}, \quad \widetilde{K}_j = \frac{\Delta t}{2} K_j + \frac{|E|}{3} \mathbf{I}$$

$$\widehat{K}_j = \frac{\Delta t}{2} K_j - \frac{|E|}{3} \mathbf{I}$$

with $\mathbf{I}$ the identity matrix.

6.1.2 Analysis of matrix $\mathcal{RD}$ schemes

Only a few comments are given on the formal properties of matrix $\mathcal{RD}$ schemes. The easiest task is the characterization of the accuracy of the discretization, as the analysis of Section 4.2.1 generalizes immediately to systems. Also in the case of the energy stability, the results available for the linear first-order schemes extend to their matrix counterpart (Abgrall and Barth, 2002; Abgrall and Barth, 2001; Abgrall and Mezine, 2003b; Barth, 1996; Abgrall and Mezine, 2004; Abgrall and Mezine, 2003a).

The issue of the nonoscillatory character of the discretization is more delicate. In the system case, a real maximum principle for the exact solution does not exist. In the framework of $\mathcal{RD}$ schemes, an attempt to give a characterization of the nonoscillatory character of discretizations of (161) is due to Abgrall and Mezine (2004) and Abgrall and Mezine (2003b). Arguing that solutions to (161) are generally piecewise smooth, with no oscillations in correspondence of discontinuities, in the references the authors use a wave decomposition technique to derive estimates on the maximum norm of the components of the nodal solution vectors for the case of symmetric systems. Here, we give no further details on this, referring to the mentioned papers for details. We limit ourselves to recalling that the matrix variants of scalar first-order LED schemes do respect a monotonicity condition, in the sense of Abgrall and Mezine (2004).

Finally, concerning the construction of nonlinear limited matrix distribution schemes (cf. Section 4.4.2), we recall the technique proposed in Abgrall and Mezine (2004), and used to obtain the results discussed later. This technique is thoroughly discussed in Abgrall and Mezine (2004) and Abgrall and Mezine (2003b) and finds its theoretical justification in the L_∞ stability criterion introduced in the same references. The idea is quite simple: given a monotone scheme with nodal residuals $\boldsymbol{\phi}_i^{\mathcal{M}}$ and a local direction $\vec{\xi}$, decompose the nodal residuals as

$$\boldsymbol{\phi}_i^{\mathcal{M}} = \sum_\sigma \langle \mathbf{l}^\sigma, \boldsymbol{\phi}_i^{\mathcal{M}} \rangle \mathbf{r}^\sigma = \sum_\sigma \varphi_i^{\mathcal{M},\sigma} \mathbf{r}^\sigma$$

with $\mathbf{l}^\sigma$ the left eigenvectors of $K = A_1 \xi_1 + A_2 \xi_2$. Each $\varphi_i^{\mathcal{M},\sigma}$ is treated as a scalar residual, and limited. The nodal residuals of the nonlinear scheme are obtained by projecting back in physical space: $\boldsymbol{\phi}_i = \sum_\sigma \varphi_i^\sigma \mathbf{r}^\sigma$. The mappings used here are basically the same that can be used in the scalar case (Abgrall and Mezine, 2004; Abgrall and Mezine, 2003b; Abgrall and Roe, 2003), for exmple, mapping (105) (or equivalently (106)). Note that, even though the contruction makes use of an *arbitrary* direction $\vec{\xi}$, in practice the results are little affected by its choice (Abgrall and Mezine, 2004).

While the well-posedness of the procedure is still subject to Proposition 10 (applied to each scalar wave), at present no results exist on the stability of the resulting nonlinear scheme, in the L^2 sense. It is observed in practice that these $\mathcal{LP}$ nonlinear schemes show a very sharp and monotone capturing of single or interacting discontinuities. However, their performances are not entirely satisfactory in smooth regions. As already remarked, this fact is a consequence of the strong L^∞ flavor of the construction, and it is still a subject of research. We refer the reader to Abgrall (2006) and Ricchiuto and Abgrall (2006) for a recent study of this issue.

6.1.3 Nonlinear systems of conservation laws

As in the scalar case, the passage to nonlinear conservation laws has to guarantee the conservative nature of the final discretization. The thorough discussion of Section 4.5.1 is also valid in the system case, and will not be repeated here. We limit ourselves to recalling an important particular case, for which a simple exact mean-value linearization exists. Then we recall more general ways of handling the nonlinearity (Abgrall and Barth, 2002; Csík, Ricchiuto and Deconinck, 2002; Ricchiuto, Csík and Deconinck, 2005).

Consider the issue of computing steady solutions to (1) in the homogeneous case. Given an unstructured discretization of the spatial domain, we proceed as in Definition 1, and compute the spatial residual

$$\boldsymbol{\phi}^E = \int_E \nabla \cdot \boldsymbol{\mathcal{F}}_h(\mathbf{u}_h) \, \mathrm{d}x \, \mathrm{d}y$$

Let us denote by $\mathbf{w}$ a set of primitive variables (not necessarily $\mathbf{u}$) that are assumed to vary piecewise linearly over the mesh, as in (9). We rewrite the residual as

$$\begin{aligned} \boldsymbol{\phi}^E &= \int_E \frac{\partial \boldsymbol{\mathcal{F}}(\mathbf{w}_h)}{\partial \mathbf{w}} \cdot \nabla \mathbf{w}_h \, \mathrm{d}x \, \mathrm{d}y \\ &= \left(\int_E \frac{\partial \boldsymbol{\mathcal{F}}(\mathbf{w}_h)}{\partial \mathbf{w}} \, \mathrm{d}x \, \mathrm{d}y \right) \cdot \nabla \mathbf{w}_h|_E = \sum_{j \in E} \widetilde{K}_j \mathbf{w}_j \end{aligned}$$

with $\boldsymbol{\mathcal{F}}(\mathbf{w}_h) = \boldsymbol{\mathcal{F}}(\mathbf{u}(\mathbf{w}_h))$, and with

$$\widetilde{K}_j = \frac{1}{2} \frac{\widetilde{\partial \boldsymbol{\mathcal{F}}}}{\partial \mathbf{w}} \cdot \vec{n}_j, \quad \frac{\widetilde{\partial \boldsymbol{\mathcal{F}}}}{\partial \mathbf{w}} = \frac{1}{|E|} \int_E \frac{\partial \boldsymbol{\mathcal{F}}(\mathbf{w}_h)}{\partial \mathbf{w}} \, \mathrm{d}x \, \mathrm{d}y \quad (169)$$

The computation of the exact mean-value flux Jacobian, needed in the definition of $\widetilde{K}_j$, can be quite costly, if not impossible. While the technique proposed in Abgrall and Barth (2002) (see also Section 4.5.1) represents a practical solution to get a good approximation of this quantity for any arbitrary (symmetrizable) system of conservation laws, the issue of the cost of the computation remains. An important exception to this is the system of the Euler equations for a perfect gas. In this case, in fact, it has been shown that the components of the flux tensor can be written as quadratic polynomials in terms of the components of a multidimensional generalization of Roe's parameter vector $\mathbf{z}$ (Roe, 1981). Hence, assuming $\mathbf{z}_h$ to be the piecewise linear variable, as in (9), and the entries of the flux Jacobians being linear in the components of $\mathbf{z}_h$, an exact mean-value linearization is obtained simply by evaluating them in the arithmetic average of the nodal values of $\mathbf{z}_h$ over E. This simple conservative linearization, due to Deconinck, Roe and Struijs (1993), allows the application of matrix $\mathcal{RD}$ schemes to the Euler equations in a simple and effective way.

The Euler equations for a perfect gas are a *lucky coincidence*, simple conservative linearizations being in general hard to find. One solution to this problem is the approach based on approximate Gaussian quadrature of Abgrall and Barth (2002), which however has the drawback of being computationally demanding. In practice, the most effective approach is the conservative formulation of Csík, Ricchiuto and Deconinck (2002) and Ricchiuto, Csík and Deconinck (2005). The elements given in this section, and in Sections 4.5.1 and 5.4, allow us to easily write down conservative matrix variants of $\mathcal{RD}$ schemes, just by replacing the scalar upwind parameters k_j^+ with matrix upwind parameters K_j^+ evaluated using an arbitrary linearized state over the element.

6.2 Some numerical results

To show the potential of the $\mathcal{RD}$ approach in computing complex solutions of nonlinear conservation laws, we present some illustrative numerical results. To validate the $\mathcal{RD}$ approach, we use well-known tests involving the solution of the Euler equations for a perfect gas. To show the flexibility of the approach adopted for conservation, we consider the solution of a simple model of homogeneous two-phase flow, which, owing to the nonlinearity of the equations of state (EOS), presents all the generality of systems of conservation laws with complex thermodynamics. Finally, the shallow-water equations are chosen as an application showing the potential of residual-based discretizations.

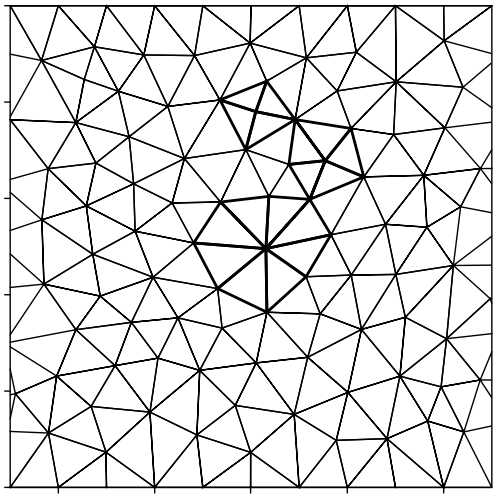

Figure 14. Unstructured triangulation.

All the results presented are obtained on grids with the *irregular* topology reported on Figure 14, using schemes based on the conservative approach of Csík, Ricchiuto and Deconinck (2002) and Ricchiuto, Csík and Deconinck (2005). The distribution is achieved either by means of the conservative matrix variant of the N scheme (154) (in the following text simply referred to as the N scheme), or its limited nonlinear $\mathcal{LP}$ variant (in the following text referred to as the *LN scheme*).

Euler equations: double Mach reflection

This problem is a severe test for the robustness and the accuracy of schemes designed to compute discontinuous flows containing complex structures (Woodward and Colella, 1984). It consists of the interaction of a planar right-moving Mach 10 shock with a 30 degree ramp. We refer to Woodward and Colella (1984) for details concerning the setup of the test. The simulation has been run on an unstructured triangulation with $h = 1/100$ until time $t_f = 0.2$. We present in Figure 15 the density contours obtained (on the same mesh) with the LN scheme and with a second-order cell-centered $\mathcal{FV}$ scheme using Roe's numerical flux, linear reconstruction and limiter proposed in Barth and Jespersen (1989), and a second-order Runge–Kutta (RK) time integrator. The $\mathcal{RD}$ LN scheme clearly shows sharper approximation of the shocks, and a much better resolution of the contact emanating from the triple point and of the jet on the ramp.

Euler equations: a shock–shock interaction

We consider one of the two-dimensional Riemann problems studied in Kurganov and Tadmor (2002) and later also in De Palma *et al.* (2005) and Ricchiuto, Csík and Deconinck (2005). It consists of the interaction of two oblique shocks with two normal shocks. See Figure 16 for a sketch of

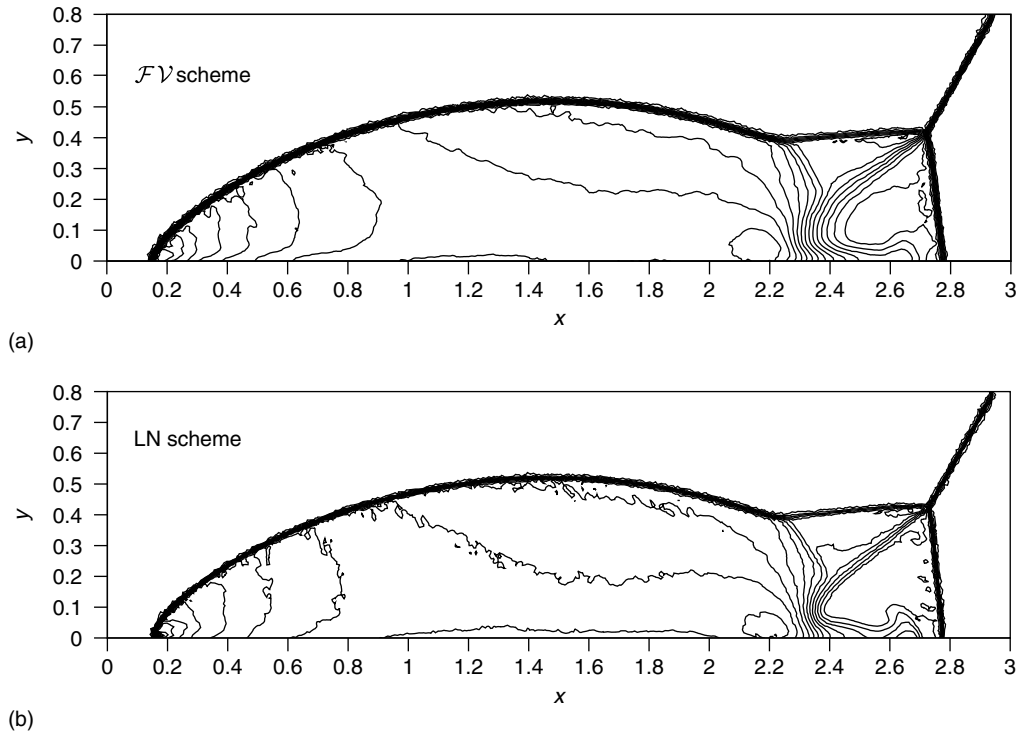

Figure 15. Double Mach reflection. Cell-centered $\mathcal{FV}$ scheme (a), and LN scheme (b).

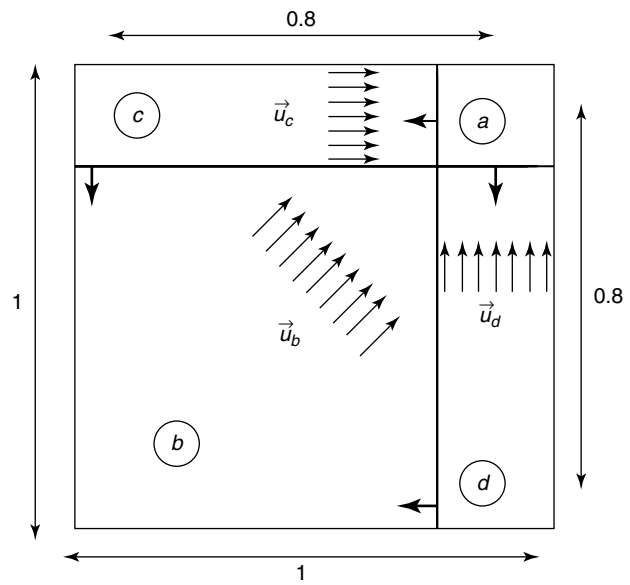

Figure 16. Shock−shock interaction. Initial solution.

the initial solution (details can be found in Kurganov and Tadmor, 2002, De Palma *et al.*, 2005, and Ricchiuto, Csík and Deconinck, 2005). We compare the results obtained on the same grid ($h = 1/200$) with the LN scheme, and with a second-order cell-centered $\mathcal{FV}$ scheme (Barth and Jespersen, 1989) with second-order RK time integration.

We visualize the contours of the density on Figure 17. The nature of the flow is quite complex. The interaction generates two symmetric lambda-shaped couples of shocks and a downward moving normal shock. Strong slip lines emanate from the lower triple points and interact with one of the branches of the upper lambda-shocks. A jet of fluid is pushed from the high-pressure region (state a in Figure 16) against the normal shock. Compared to the $\mathcal{FV}$ scheme, the LN scheme gives a richer solution. The onset of Kevin-Helmholtz instabilities along the contact lines interacting with the upper lambda-shock is already visible on this mesh.

6.2.1 A two-phase flow model

Consider now the system of conservation laws defined by the following set of conserved variables and fluxes:

$$\mathbf{u} = \begin{bmatrix} \alpha_g \rho_g \\ \alpha_l \rho_l \\ \rho u \\ \rho v \end{bmatrix}, \quad \boldsymbol{\mathcal{F}}(\mathbf{u}) = \begin{bmatrix} \alpha_g \rho_g u & \alpha_g \rho_g v \\ \alpha_l \rho_l u & \alpha_l \rho_l v \\ \rho u^2 + p & \rho u v \\ \rho u v & \rho v^2 + p \end{bmatrix} \tag{170}$$

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

$$\rho = \alpha_g \rho_g + \alpha_l \rho_l \tag{171}$$

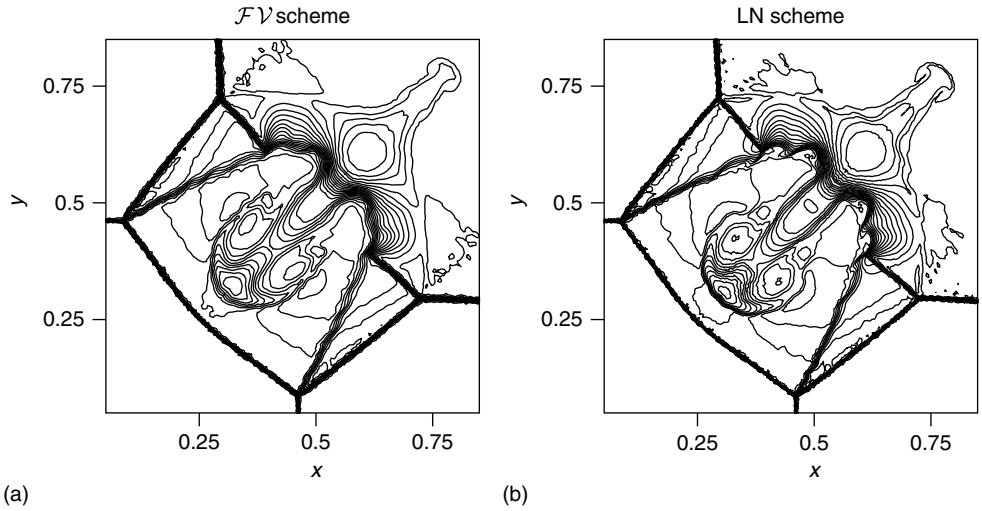

Figure 17. Shock–shock interaction. $\mathcal{FV}$ scheme (a) and LN scheme (b).

and p is the pressure. The model is closed by the relation $\alpha_g + \alpha_l = 1$, and by the EOS relating the densities to the pressure. In the following text, we will denote by α the gas volume fraction, often referred to as the *void* fraction. We assume implicitly that $\alpha_l = 1 - \alpha$. Concerning the EOS, we have used, as in Paillère, Corre and Garcia (2003), the following relations representative of air and water:

$$p = \Gamma_g \left(\frac{\rho_g}{\rho_{g0}}\right)^{\gamma_g} \quad \text{and} \quad p = \Gamma_l \left[\left(\frac{\rho_l}{\rho_{l0}}\right)^{\gamma_l} - 1\right] + p_{l0} \tag{172}$$

The values of all the constants in the EOS are taken as in Paillère, Corre and Garcia (2003), Ricchiuto, Csík and Deconinck (2005), and Ricchiuto (2005). This system of conservation laws constitutes a fairly simple model of homogeneous air-water two-phase flow. However, the relation between the pressure and the conserved mass and momentum fluxes is so complex that a conservative linearization can hardly be derived. Because of the nonlinearity of the EOS, pressure and volume fractions cannot be computed in closed form from the conserved variables. Instead, combining the EOS and the relation $\alpha_l = 1 - \alpha$, a nonlinear equation for the pressure is obtained, which can be solved in a few Newton iterations (Paillère, Corre and Garcia, 2003).

Two-phase flow model: Mach 3 moving shock

We consider 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 $[0, 2] \times [0, 0.1]$. We have run the simulations on an irregular triangulation with element size $h = 1/100$. Periodic boundary conditions are imposed on the top and bottom boundaries. The final time of the simulation corresponds to a displacement of the *exact* shock location (computed analytically) by a unit length. We present the solutions of the conservative N and LN schemes. The output is visualized by extracting the data along the line $y = 0.05$. The results are reported in Figure 18. The shock position is correctly simulated, confirming the conservative character of the discretization. The nonoscillatory character of the results is evident. The nonlinear scheme gives a very sharp and monotone capturing of the discontinuity.

Two-phase flow model: a shock-bubble interaction

We now consider a two-phase variant of the shock-bubble interaction presented for the Euler equations. The initial solution 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 $r_b = 0.2$. We present the results obtained with the LN and LST-N schemes on an unstructured grid with reference element size $h = 1/200$, at several time instances.

From the Figure 19 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. Once the *undisturbed* shock has crossed the region occupied by the whole circular discontinuity, and has joined the transmitted shock, the interface of the contact folds, rolling-up into a symmetric structure. The LN

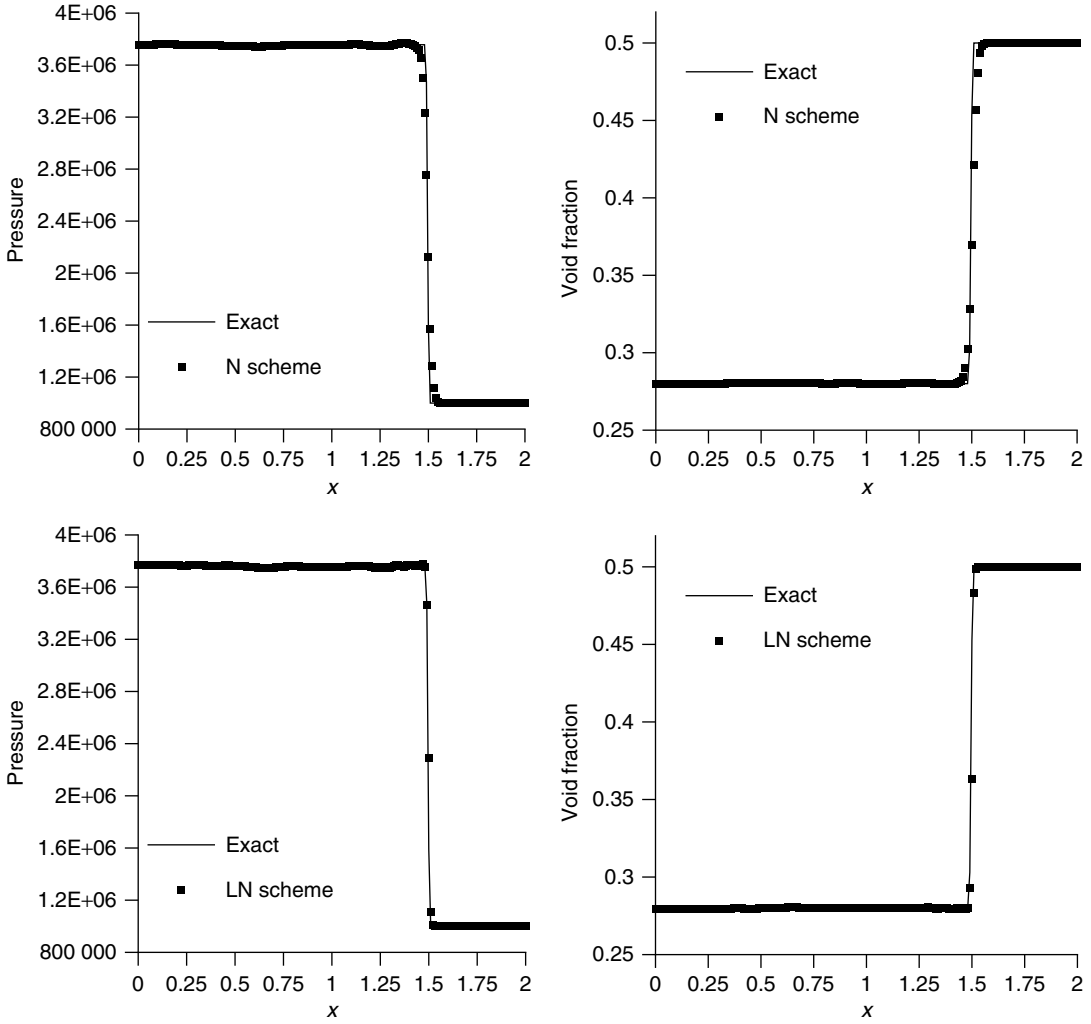

Figure 18. Two-phase $M_S = 3$ shock. Pressure (left) and void fraction (right) along the line $y = 0.05$. Solutions of the N (top) and LN (bottom) schemes.

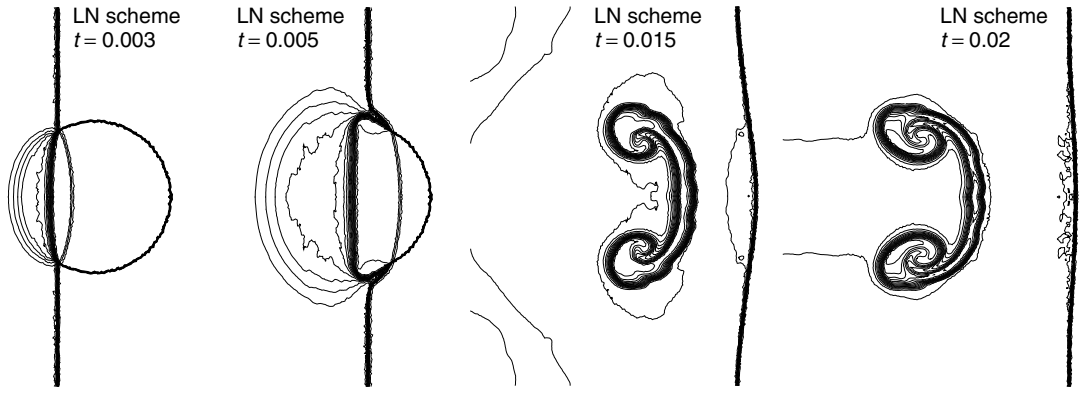

Figure 19. Two-phase shock-bubble interaction, LN scheme. Mixture density at $t = 0.003$, $t = 0.005$, $t = 0.015$, and $t = 0.02$.

scheme gives a crisp resolution of the contact, its wavy structure showing the glimpse of an inviscid instability.

6.2.2 *Shallow water flows: $\mathcal{RD}$ schemes and well-balancedness*

This last section discusses a few of the results obtained in Ricchiuto (2005) and Ricchiuto, Abgrall and Deconinck (2007) by solving the shallow-water equations with conservative schemes of Csík, Ricchiuto and Deconinck (2002) and Ricchiuto, Csík and Deconinck (2005). This system of equations can be written as (1), with

$$\mathbf{u} = \begin{bmatrix} H \\ Hu \\ Hv \end{bmatrix}, \quad \boldsymbol{\mathcal{F}}(\mathbf{u}) = \begin{bmatrix} Hu & Hv \\ Hu^2 + g\dfrac{H^2}{2} & Huv \\ Huv & Hv^2 + g\dfrac{H^2}{2} \end{bmatrix}$$

$$\boldsymbol{\mathcal{S}}(\mathbf{u}, x, y) = -gH \begin{bmatrix} 0 \\ \dfrac{\partial B(x,y)}{\partial x} \\ \dfrac{\partial B(x,y)}{\partial y} \end{bmatrix} \tag{173}$$

with H the *local relative water height*, $\vec{u} = (u, v)$ the flow speed, g the gravity acceleration, and $B(x, y)$ the local height of the bottom (see Figure 20). We also define the total water height $H_{\text{tot}} = H + B$.

The shallow-water equations admit several classes of known exact solutions. Among these, we are interested in the lake-at-rest solution

$$H_{\text{tot}} = H_0 = const, \quad H = H_0 - B(x, y), \quad u = v = 0$$

The following result is proved in Ricchiuto (2005) and Ricchiuto, Abgrall and Deconinck (2007).

Proposition 16 ($\mathcal{LP}$ schemes and the lake-at-rest solution) *$\mathcal{LP}\,\mathcal{RD}$ schemes preserve exactly the lake-at-rest solution, provided that the same numerical approximation is used for H and B, and provided that the element residual is computed exactly with respect to this approximation. This is true independently on topology of the mesh, the regularity of $B(x, y)$, and polynomial degree of the approximation.*

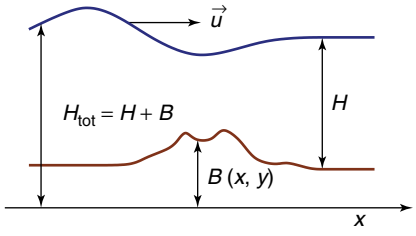

Figure 20. Shallow-water equations: basic unknowns.

This proposition shows the big advantage of the residual approach at the basis of the $\mathcal{RD}$ discretization, and generalizes the numerical observations of Brufau and Garcia-Navarro (2003) and references therein.

Still flow over smooth bed

We verify Proposition 16 experimentally. On the domain $[0, 1]^2$, we consider an initial state in which the velocity is zero and $H = 1 - B(x, y)$ with (LeVeque, 1998; Xing and Shu, 2005; Seaïd, 2004)

$$B(x, y) = 0.8e^{-50((x-0.5)^2+(y-0.5)^2)}$$

We compute the solution up to time $t = 0.5$ with the LN scheme on an irregular triangulation and $h = 1/100$. In Table 1, we report the values (computation run in double precision) of the norms of the errors on water height and velocity components. The results obviously confirm the theoretical result of the proposition. The numerical output is similar $\forall\, t > 0$.

Water height perturbation over smooth bed

We now consider a problem involving a perturbation of the exact lake-at-rest solution. The objective is to verify that the schemes are able to resolve the evolution of the perturbation and its interaction with the bed shape, without spoiling the exact lake-at-rest state in unperturbed regions. The spatial domain is the rectangle $[0, 2] \times [0, 1]$. Initial state and bottom shape are chosen as in Seaïd (2004) and Xing and Shu (2005). In particular, we set

$$B(x, y) = 0.8e^{-5(x-0.9)^2-50(y-0.5)^2}$$

At $t = 0$ the velocity is set to zero everywhere, while the relative water height is set to

$$H = \begin{cases} 1.01 - B(x, y) & \text{if} \quad 0.05 < x < 0.15 \\ 1 - B(x, y) & \text{otherwise} \end{cases}$$

We solve the problem on an unstructured discretization of the domain with reference mesh size $h = 1/100$. In Figure 21 we visualize the results obtained with the space-time LN scheme at $t = 0.12$ and $t = 0.48$, in terms of

Table 1. Norm of the errors at time $t = 0.5$, LN scheme.

	L^∞	L^1	L^2
H	7.491837e-17	7.085969e-17	7.107835e-17
u	7.478237e-17	7.161000e-17	7.169336e-17
v	7.478237e-17	7.177553e-17	7.177653e-17

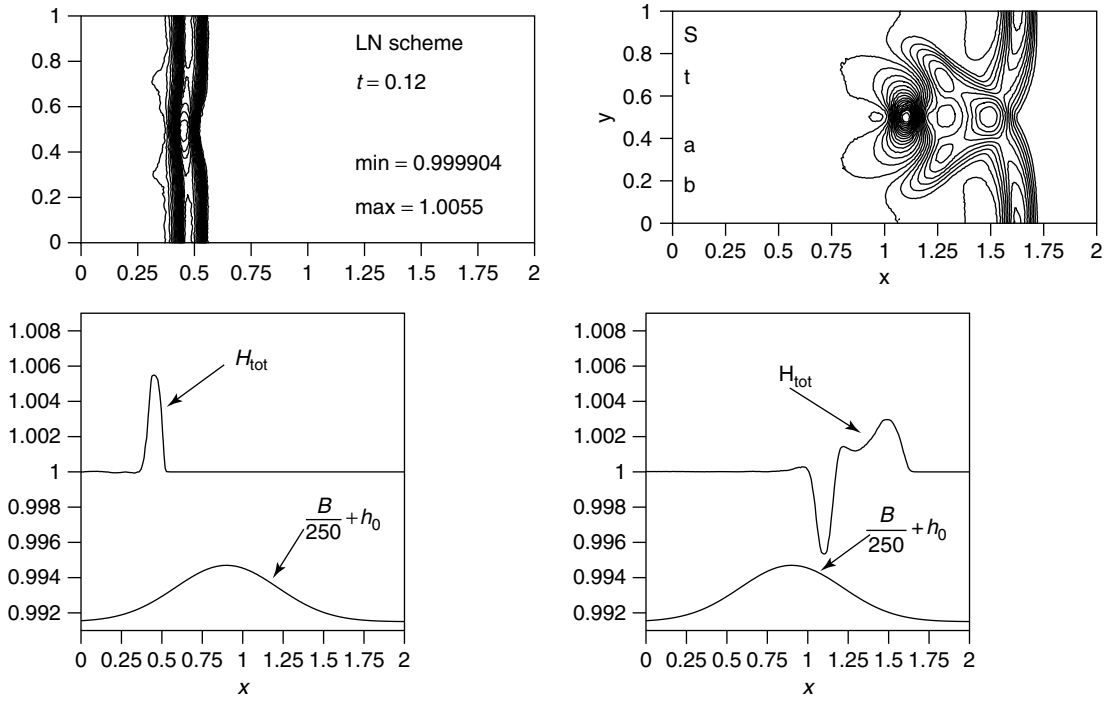

Figure 21. Water height perturbation over smooth bed. Solution of the LN scheme at $t = 0.12$ (left) and $t = 0.48$ (right). Top: contour plot of total water height. Bottom: distribution of H_{tot} at $y = 0.5$ ($h_0 = 0.9915$).

total water height contours (top pictures) and water height distribution at $y = 0.5$ (bottom pictures). The following observations can be made. In the region ahead of the perturbation, the exact solution is *perfectly preserved* up to machine accuracy, while behind the perturbation the solution quickly gets back to the lake-at-rest state. Compared to the results of Xing and Shu (2005), obtained with a fifth-order finite difference weighted essentially non-oscillatory (WENO) scheme on a *structured mesh* with $h = 1/100$, our results reproduce the interaction well. The small structures contained in the reference solution are visible in the results of the LN scheme. Obviously, the use of a very high-order discretization is beneficial when approximating this type of problem. Nevertheless, on an irregular unstructured triangulation, the LN scheme gives a rich solution structure, while yielding a monotone approximation, and preserving exactly the lake-at-rest state in the unperturbed region.

7 CONCLUSIONS, ONGOING WORK, AND OPEN ISSUES

7.1 General summary

This paper has reviewed the basics of the $\mathcal{RD}$ methodology and some of its theoretical foundations. We have tried to analyze the core ingredients of the method rather than focusing on engineering applications. These ingredients are as follows:

1. a simple way to obtain high-order accuracy on unstructured grids, depending only on the polynomial approximation space and on the use of a bounded distribution strategy;
2. the strong emphasis on L_∞ stabilization (viz. monotonicity) relying on the theory of positive coefficient schemes;
3. a truly multidimensional upwinding strategy allowing to more faithfully mimic the directional propagation of the information typical of solutions to conservation laws;
4. the extension to time-dependent equations by a space-time approach;
5. a general conservative formulation that does not rely on any ad-hoc linearization;
6. the extension to systems by an algebraic matrix generalization justified by a simple wave analysis.

In no way this contribution has given an exhaustive overview of the present state of the method and of the ongoing research on the subject of residual-based discretizations. For example, we have completely left out the recent work of Lerat and Corre on RBC (residual-based compact) schemes (Lerat and Corre, 2001; Lerat and Corre, 2002; Corre, Hanss and Lerat, 2005), which is based on the very

same idea of constructing a discretization in which the main actor is a local approximation of the original mathematical equation as a whole, rather than a discrete approximation of the partial derivative themselves.

Despite the effort put in its development since the 1980s, it is still not possible to claim that the method is mature, although impressive results have been obtained. The domain is still in full development and much progress can be expected in the coming years. In the following paragraphs, we attempt to illustrate the presently most active research lines, by reviewing recent work of the different groups active in this field. We also try to underline the weak points, and of course the areas for future research.

7.2 Monotonicity and stability

Owing to their nature, $\mathcal{RD}$ schemes hardly allow a proper stability analysis, especially in the energy or entropy norms. The understanding of the importance of multidimensional upwinding itself in this matter is probably quite limited. This partly explains the fact that for years most (or all) of the nonlinear $\mathcal{RD}$ discretizations that have been proposed in literature suffer from a lack of nonlinear iterative convergence, often endangering the ability of obtaining optimal grid convergence rates, especially when dealing with nonlinear systems.

This has been found to be true for most blended schemes (see Section 4.4.1 and Abgrall, 2001; Csík, Ricchiuto and Deconinck, 2002; Dobeš and Deconinck, 2006), and it is especially true for the limited schemes for systems, briefly described in Section 6.1.2. An analysis of this problem has been given recently in Abgrall (2006). In the reference, it is argued that the problem is of algebraic nature, and it is strictly tied to the issue of properly defining upwinding, especially for systems. The cure proposed in the reference is based on a stabilization technique since long known in the finite-element community, and very close to the one used in Lerat and Corre (2001), Lerat and Corre (2002), and Corre, Hanss and Lerat (2005) to construct dissipative RBC schemes.

We also mention that different strategies for constructing nonoscillatory $\mathcal{RD}$ discretizations have been proposed. Among these, the flux corrected transport (FCT) procedure of Zalesak (1979), has been used for example in Hubbard and Roe (2000) and De Palma, Pascazio and Napolitano (2001). A novel procedure, with many similarities with FCT, has also been recently proposed in De Palma *et al.* (2005). $\mathcal{RD}$ schemes based on WENO reconstructions have also recently appeared (Chou and Shu, 2006), while, in the context of higher-order schemes, the idea of limiting the polynomial representation of the variables (in a more classical, $\mathcal{FV}$-like sense), has been put forward in Hubbard (2007).

At present, a throughout comparison of these different approaches has not been performed. Certainly, the study of nonlinear nonoscillatory discretizations is far from finished, almost each group proposing a different strategy. Perhaps, better schemes will come also with better understanding of the dissipative properties of $\mathcal{RD}$ in the system case. This remains, however, one of the important open issues concerning $\mathcal{RD}$; especially, in general situations (item-dependent, source terms, very high order of accuracy, etc.).

7.3 Very high order of accuracy

One of the most basic issues left out of this paper is the construction of schemes of arbitrary high order of accuracy. The formal theoretical framework to achieve this is mainly due to Abgrall and Roe (2003) (see also Abgrall, 2005). We also refer to the analysis reported in Ricchiuto, Abgrall and Deconinck (2007), which is inspiration for Section 3.2. The potential of $\mathcal{RD}$ discretizations in this sense is summarized by Definition 3: as long as the splitting is performed with bounded distribution coefficients, the formal accuracy is determined only by the polynomial representation of the unknown used to compute the element residual. This observation is the basis for almost all the proposed higher-order discretization of the $\mathcal{RD}$ type. Concerning the respect to the improved polynomial representation, two basic approaches have appeared in literature.

The first, originally proposed in Caraeni (2000) (see also Caraeni and Fuchs, 2005), is based on a reconstruction of the gradient of the solution, which of course allows a local higher-order polynomial approximation. Developments on this line have been proposed by several authors. We mention the work of Nishikawa, Rad and Roe (2001) where this strategy has been coupled to a clever splitting of the hyperbolic and elliptic parts of the two-dimensional Euler equations, each solved with a higher-order scheme (upwind for the hyperbolic equations, while a least squares type approach is used for the elliptic operator). The overall strategy allows to compute very accurately solutions in flow regimes ranging from potential to supersonic flow. We also mention the recent work of Rossiello *et al.* (2007), where improved formulations of the third-order schemes of Caraeni (2000) are proposed. Following the analysis of Abgrall and Roe (2003) and Abgrall (2005), in Rossiello *et al.* (2007) the authors derive the conditions for a scheme to be kth order, and then propose schemes with improved stability with respect to the ones of Caraeni (2000). Monotonicity is enforced via the procedure proposed by the same authors in De Palma *et al.* (2005).

A different framework has been set up instead in Abgrall and Roe (2003). The idea in this case is to locally store all the DOF necessary to build a second (or higher) degree polynomial approximation of the unknowns, and derive discrete equations for all the DOF. In its original formulation, the method employs P^k Lagrangian triangular finite elements, k being the degree of the discrete polynomial, even though any other continuous set of polynomial basis functions can be chosen. Unpublished results for a steady-state convection equation using the LDA scheme on subtriangulated P2 triangles were already obtained in 1995 by the first author and T.J. Barth during a summer visit at NASA Ames. Discretizations built following this philosophy have been proposed for example in Abgrall and Tavé (2006), where following Abgrall (2006) a procedure to construct simplified stable and monotone centered very high order schemes is described, and in Ricchiuto *et al.* (2005) where upwind $\mathcal{RD}$ schemes up to fourth order are presented. We also mention the work of Ricchiuto, Abgrall and Deconinck (2003) and Abgrall, Andrianov and Mezine (2005), in which the same approach is used to construct schemes for time-dependent problems.

These two basic philosophies have been compared in Hubbard (2007), where a novel strategy to build monotone schemes via an edge-based limitation of the unknown polynomial variation is proposed, and in Nishikawa (2005), where several ways of including higher (second) order derivative terms in the discretization are discussed.

It is worth mentioning that some work has also been done in a different direction, namely trying to improve the accuracy by extending the stencil of the schemes. At present, this technique only works on structured, and non-smooth structured meshes. We mention, as an example, the schemes proposed in Hubbard and Laird (2005), in which a truncation error analysis has been used to devise extended stencil distribution strategies on structured grids, allowing to reach third order of accuracy. In Chou and Shu (2006), instead, the idea of k-exactness preservation of Definition 3 has been combined with a (nonlocal) WENO variable representation to build schemes up to fourth order of accuracy. This work has been extended to viscous problems in Chou and Shu (2007).

Two possible directions for future research can be mentioned: the study of improved polynomial approximations and the understanding of the discrete stability of higher-order residual distribution schemes. Regarding the polynomial approximation of the unknown, at present two main approaches exist: the reconstruction of solution gradients at the nodes to locally build a higher-degree polynomial, and the use of higher-degree Lagrange finite elements (see Hubbard, 2007 and references therein for a review). Possible alternatives could be the use of ENO/WENO approximations, as in Chou and Shu, 2006, or the use of finite elements based polynomial interpolants through Gauss-Lobatto points (see for example Eskilsonn and Sherwin, 2005 and references therein). As far as the stability of higher-order residual distribution schemes is concerned, an important point will be the understanding of the conditions to be verified for the discrete equations to admit a unique solution. The need for a stability criterion is evident from the work reported in many recent publications (Abgrall and Tavé, 2006; Chou and Shu, 2007; Rossiello *et al.*, 2007). The lack of a general variational formulation allowing the construction of energy estimates or of any other theoretical tool will certainly make the derivation of such conditions a difficult task.

7.4 Quadrilaterals and hybrid meshes

Another basic issue, which is not dealt with in this contribution, is the use of the $\mathcal{RD}$ idea on nontriangular meshes. Two main difficulties arise when trying to perform such an extension.

The first is related to conservation and to the lack of a simple conservative linearization on general elements. This issue is analyzed in detail in Abgrall and Barth (2002). However, the most successfully approach to deal with this problem appears to be the one proposed in Quintino *et al.* (2002) (see also Csík, Ricchiuto and Deconinck, 2002), later adopted by all authors in the field.

The second difficulty is peculiar to the case of quadrilateral (hexahedral in three dimensions) elements, and continuous variable representation. It is in fact long known (Rudgyard, 1993) that on quadrilateral meshes $\mathcal{LP}$ schemes suffer from the appearance of spurious modes polluting the numerical results. On regular quadrilateral grids, this has been shown with a Fourier analysis in Rubino (2006) and De Palma *et al.* (2006). For second-order schemes, the Fourier analysis nicely allows to highlight the high-frequency instability flawing all $\mathcal{LP}$ schemes (and in general k-exact schemes, with bounded distribution coefficients); we refer to Rubino (2006) for the general analysis. Surprisingly enough, one way to cure this instability comes from the work of Abgrall (2006), concerning the analysis of nonlinear limited schemes. Even though different in nature, the instability on quadrilaterals can be suppressed by adding to the discretization the same dissipation term used in the reference to stabilize the nonlinear schemes. This has been clearly shown in Abgrall and Marpeau (2007) for the scalar and system case. In the scalar case, the same technique has been used in De Palma *et al.* (2006).

With respect to the same subject, it is worth mentioning that the WENO $\mathcal{RD}$ schemes of Chou and Shu

(2007) operate on quadrilateral meshes. Being based on a discontinuous WENO approximation of the unknown, however, the schemes proposed in these references do not seem to suffer from the same type of instability.

7.5 Time-dependent problems

The improvement of $\mathcal{RD}$ discretization for unsteady problems is also a very important subject of research. It is now evident that the way to go is to define an element residual containing the time derivative. This can however be done in different ways.

One way to do this is to investigate on the consistent form of the mass matrix. As already mentioned, an interesting analysis on this subject, based on geometrical reasoning, is performed in De Palma *et al.* (2005). In the reference, conditions allowing construction of consistent mass matrices for second-order schemes are given. Following the ideas of Caraeni (2000), an attempt to extend this work to third order of accuracy can be found in Rossiello *et al.* (2007).

A different line of research considers the use of simplified formulations, following the ideas proposed in Abgrall (2006). Promising initial results have been presented in Ricchiuto and Abgrall (2006) and Bollermann (2006).

Finally, concerning the construction of very high order schemes for time-dependent problems, we also mention the work of Abgrall, Andrianov and Mezine (2005), and Ricchiuto, Abgrall and Deconinck (2003).

Also in this case, the greatest challenge is represented by the understanding of the stability of the discretization. From this point of view, space-time schemes might present some advantages, especially when going to very high degree polynomial interpolation.

7.6 Viscous flow

The extension of the $\mathcal{RD}$ philosophy to the solution of viscous flow problems seems to be one of the hardest problems at present. Even for continuous variable representations, the gradient of the discrete unknowns is always discontinuous across element edges, which renders a straightforward extension of the schemes impossible. To deal with this fact, two different approaches have appeared in literature.

The first approach has its roots in the initial work of Caraeni (2000) in which the author tries to exploit at best the properties of $\mathcal{LP}$ schemes by defining a residual that contains the second-order derivatives as well. To deal with the discontinuity of the gradient of the discrete unknowns (hence of the viscous fluxes), the author proposes to perform a reconstruction of the nodal gradients of the variable, which are then used to build a locally continuous polynomial for the gradient. The main works that have followed these lines are the ones of Nishikawa and Roe (2004), Nishikawa and Roe (2005b), Nishikawa (2005), and Nishikawa and Roe (2006). The idea emerging from the references is that, by rewriting an advective-diffusive problem as a first-order system, one can then apply the $\mathcal{RD}$ technology without any complication concerning the variable representation. Since the gradient of the solution is part of the discrete unknowns, and hence represented with the same continuous polynomial, a residual containing the dissipative fluxes can be easily defined. The main drawback of this approach is perhaps the augmentation of the number of unknowns, since now the components of the viscous fluxes (more generally the components of the viscous stress tensor) have to be solved for explicitly. Note however, that a simple variable count will still show a net advantage over methods based on discontinuous variable representation such as discontinuous Galerkin.

A different approach tries to make use of a PG analogy to couple the $\mathcal{RD}$ discrete advective operator, with a Galerkin or PG discrete approximation of the diffusion operator. In this approach, the main problem is to properly define the coupling between the $\mathcal{RD}$ discretization of the advective operator, with the Galerkin, or PG, approximation of the diffusion term, such that the overall discretization has uniform accuracy over the whole range of cell Reynolds numbers (i.e. mesh sizes). The reader may consult Ricchiuto *et al.* (2005) and Ricchiuto *et al.* (2007) for an overview on how this can be achieved. This approach does not introduce extra variables. The problem is that no unique PG formulation of $\mathcal{RD}$ exists, and that a sound variational formulation allowing stability and error analysis is also lacking. As a consequence, a real understanding of the properties of the discretization is hard to achieve, and choices are often made by intuition and analogy with classical PG discretizations.

It is worth mentioning a Fourier analysis of discretizations for the advection diffusion equation performed in De Palma *et al.* (2006). The reader is also referred to Chou and Shu (2007) for the extension of the work done in Chou and Shu (2006) on WENO $\mathcal{RD}$ schemes to the viscous case. In the last reference, in particular, the discontinuity of the viscous fluxes posed no particular problems since the unknown itself is discontinuous, and numerical fluxes replace the physical ones at cell interfaces.

7.7 Variable representation and adaptive strategies

An interesting topic of research is also the extension of the $\mathcal{RD}$ idea to the case of discontinuous variable representation. This would allow, at least in principle, an

easier formulation of $h - p$ adaptive strategies in the $\mathcal{RD}$ framework.

At present, the only schemes of this type are the ones proposed by Chou and Shu (2006). Other researchers have however developed their versions of discontinuous $\mathcal{RD}$ schemes, aven though few publications have appeared yet on the subject (Abgrall, personal communication) (Hubbard, 2007).

Concerning adaptive strategies, we mention the work of Roe and Nishikawa (2002), allowing a proper adaptation of the mesh via a least squares minimization of the element residuals, and the adaptive quadrature proposed in Nishikawa and Roe (2005a), allowing, through a nonlinear wave detection mechanism, to avoid nonphysical expansion shocks (see also the related work of Sermeus and Deconinck, 2005 on this subject).

7.8 Applications

Engineering applications of $\mathcal{RD}$ schemes have started appearing in the most recent years. The use of $\mathcal{RD}$ schemes for turbomachinery applications has been shown in Henriques and Gato (2004), Henriques and Gato (2002) and (Bonfiglioli *et al.*, 2005). Even more complex applications including large eddy simulation (LES) simulations can be found in Caraeni (2000), and in the references therein by the same author.

Other industrial applications have started to appear, as for example in Wu and Bogy (2001), where a $\mathcal{FS}$ discretization has been used to discretize the convective term of a mathematical model used in hard disk manufacturing.

In aeronautics, we refer to the works of Edwin van der Weide and Kurt Sermeus, who developed a 3D Navier–Stokes solver including Reynolds-averaged Navier–Stokes (RANS) turbulence modeling for aeronautical applications, under the support of the European Space Agency (ESA) and the European Union 6th Framework Program Industrial demonstration of accurate and efficient multidimensional upwind and multigrid algorithms for aerodynamic simulation on unstructured grids (Project (IDEMAS)). A review of this work is given in Deconinck, Sermeus and Abgrall (2000). Extension to hypersonic applications with flow under thermal and chemical nonequilibrium had already begun in Degrez and van der Weide (1999), with 3D applications and advanced modeling accomplished over the last year. Unsteady aeronautical applications on moving geometries using an arbitrary Lagrangian–Eulerian (ALE) formulation have been developed in the PhD research thesis of Dobes (2007) with application to fluid-structure interaction.

We mention the successful application of the schemes to the solution of the shallow-water equations (Paillère, Degrez and Deconinck, 1998; Hubbard and Baines, 1997; Ricchiuto, Abgrall and Deconinck, 2007). Some of the results of the last reference have been reported in the results section of this paper. Successful extension of this work to the computation of flows with dry areas can be found in Bollermann (2006).

Finally, $\mathcal{RD}$ techniques have been also applied for magnetohydrodynamics simulations in Csík, Deconinck and Poedts (2001), Csík, Ricchiuto and Deconinck (2002), Aslan (1999), and Aslan (2004).

ACKNOWLEDGMENTS

The authors acknowledge the contributions of (in alphabetical order) R. Abgrall, T.J. Barth, D. Caraeni, M. Hubbard, M. Napolitano, P.L. Roe, M. Rudgyard, D. Sidilkover, C.-W. Shu and all their collaborators for many of the ideas discussed in this paper. The contribution of the second author was performed as a member of the doctoral program at the von Karman Institute (VKI). The authors especially acknowledge all the colleagues at VKI, too many to enumerate, with whom they worked side by side over the last years. All these people continuously helped to provide a stimulating research environment.

REFERENCES

Abgrall R. Toward the ultimate conservative scheme: following the quest. *J. Comput. Phys.* 2001; **167**(2):277–315.

Abgrall R. Very high order residual distribution methods. *VKI LS, 34*rd *Computational Fluid Dynamics Course*. von Karman Institute for Fluid Dynamics, 2005.

Abgrall R. Essentially non oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys.* 2006; **214**(2):773–808.

Abgrall R, Andrianov N and Mezine M. Towards very high-order accurate schemes for unsteady convection problems on unstructured meshes. *Int. J. Numer. Methods Fluids* 2005; **47**(8–9):679–691.

Abgrall R and Barth TJ. New results for residual distribution schemes. In *Godunov Methods. Theory and Applications*, Toro EF (ed.) International conference, Oxford, GB, October 1999. Kluwer Academic/ Plenum Publishers: New York, 2001; 27–43.

Abgrall R and Barth TJ. Residual distribution schemes for conservation laws via adaptive quadrature. *SIAM J. Sci. Comput.* 2002; **24**(3):732–769.

Abgrall R and Marpeau F. Residual distribution schemes on quadrilateral meshes. J. Sci. Comput. 2007; **30**(1):131–175.

Abgrall R, Mer K and Nkonga B. A Lax–Wendroff type theorem for residual schemes. In *Innovative Methods for Numerical*

Solutions of Partial Differential Equations, Haffez M and Chattot JJ (eds). World Scientific, 2002; 243–266.

Abgrall R and Mezine M. Construction of second-order accurate monotone and stable residual distribution schemes for unsteady flow problems. *J. Comput. Phys.* 2003a; **188**:16–55.

Abgrall R and Mezine M. Residual distribution schemes for steady problems. *VKI LS 2003–05, 33*rd *Computational Fluid Dynamics Course*. von Karman Institute for Fluid Dynamics, 2003b.

Abgrall R and Mezine M. Construction of second-order accurate monotone and stable residual distribution schemes for steady flow problems. *J. Comput. Phys.* 2004; **195**:474–507.

Abgrall R and Roe PL. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.* 2003; **19**(3):3–36.

Abgrall R and Tavé C. Construction of high order residual distribution schemes. *ICCFD4 Proceedings*. Springer-Verlag, 2006.

Aslan N. MHD-A: a fluctuation splitting wave model for planar magnetohydrodynamics. *J. Comput. Phys.* 1999; **153**(2): 437–466.

Aslan N. A visual fluctuation splitting scheme for magnetohydrodynamics with a new sonic fix and Euler limit. *J. Comput. Phys.* 2004; **197**(1):1–27.

Atkins HL and Shu C-W. *Quadrature-Free Implementation of Discontinuous Galerkin Method for Hyperbolic Equations*. Technical Report TR-96-51, ICASE, 1996.

Barth TJ. *An Energy Look at the N Scheme*. Working notes, NASA Ames Research Center: California, 1996.

Barth TJ. Numerical methods for gasdynamic systems on unstructured meshes. In *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, Vol.5 of *Lecture Notes in Computational Science and Engineering*, Kröner D, Ohlberger M and Rohde C(eds). Springer-Verlag: Heidelberg, 1998; 195–285.

Barth TJ. Numerical methods for conservation laws on structured and unstructured meshes. *VKI LS 2003-05, 33*rd *Computational Fluid Dynamics Course*. von Karman Institute for Fluid Dynamics, 2003.

Barth TJ and Jespersen DC. The design and application of upwind schemes on unstructured meshes. AIAA paper 89–0355. In *27th AIAA Aerospace Sciences Meeting*, Reno, January 1989.

Barth TJ and Ohlberger M. Finite volume methods: foundation and analysis. In *Encyclopedia of Computational Mechanics*, Stein E, de Borst R and Hughes TJR (eds). John Wiley & Sons, 2004.

Berman A and Plemmons RJ. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

Bollermann A. *On the Application of Conservative Residual Distribution Schemes to the Solution of the Shallow Water Equations on Dry Beds*. Master's thesis, IGPM, RWTH Aachen University, 2006.

Bolley C and Crouzeix M. Conservation de la positivité lors de la discétization des problèmes d'èvolution paraboliques. *Rairo Anal. Numer.* 1978; **12**:237–254.

Bonfiglioli A, De Palma P, Pascazio G and Napolitano M. An implicit fluctuation splitting scheme for turbomachinery flows. *ASME J. Turbomach.* 2005; **127**(2):395–401.

Brufau P and Garcia-Navarro P. Unsteady free surface flow simulation over complex topography with a multidimensional upwind technique. *J. Comp. Phys.* 2003; **186**(2):503–526.

Caraeni DA. *Development of a Multidimensional Upwind Residual Distribution Solver for Large Eddy Simulation of Industrial Turbulent Flows*. PhD thesis, Lund Institute of Technology, 2000.

Caraeni D and Fuchs L. Compact third–order multidimensional upwind scheme for Navier-Stokes simulations. *Theor. Comp. Fluid Dyn.* 2002; **15**:373–401.

Caraeni D and Fuchs L. Compact third-order multidimensional upwind discretization for steady and unsteady flow simulations. *Comput. Fluids* 2005; **34**(4–5):419–441.

Carette J-C, Deconinck H, Paillère H and Roe PL. Multidimensional upwinding: its relation to finite elements. *Int. J. Numer. Methods Fluids* 1995; **20**:935–955.

Chou C-S and Shu C-W. High order residual distribution conservative finite difference WENO schemes for steady state problems on non-smooth meshes. *J. Comp. Phys.* 2006; **214**(3):698–724.

Chou C-S and Shu C-W. High order residual distribution conservative finite difference WENO schemes for convection-diffusion steady state problems on non-smooth meshes. *J. Comp. Phys.* 2007; **224**(2):992–1020.

Corre C, Hanss G and Lerat A. A residual-based compact scheme for the unsteady compressible Navier-Stokes equations. *Comput. Fluids* 2005; **34**(4–5):561–580.

Csík A and Deconinck H. Space time residual distribution schemes for hyperbolic conservation laws on unstructured linear finite elements. *Int. J. Numer. Methods Fluids* 2002; **40**:573–581.

Csík A, Deconinck H and Poedts S. Monotone residual distribution schemes for the ideal magnetohydrodynamics equations on unstructured grids. *AIAA J.* 2001; **39**(8):1532–1541.

Csík A, Ricchiuto M and Deconinck H. A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws. *J. Comput. Phys.* 2002; **179**(2):286–312.

Csík A, Ricchiuto M and Deconinck H. Space time residual distribution schemes for hyperbolic conservation laws over linear and bilinear elements. *VKI LS 2003–05, 33*rd *Computational Fluid Dynamics Course*. von Karman Institute for Fluid Dynamics,2003a.

Csík A, Ricchiuto M and Deconinck H. Space-time residual distribution schemes for two-dimensional Euler and two-phase flow simulations. In *Fluid Dynamics and Aeronautics New Challenges*, *A Series of Handbooks on Theory and Engineering Applications of Computational Methods*, Priaux J, Champion M, Gagnepain J-J, Pironneau O, Stoufflet B and Thomas Ph (eds). CIMNE Barcelona, 2003b.

Deconinck H, Hirsch Ch and Peuteman J. *Characteristic Decomposition Methods for the Multidimensional Euler Equations*, Vol.264 of Lecture Notes in Physics. Springer-Verlag, 1986; 216–221.

Deconinck H, Roe PL and Struijs R. A multidimensional generalization of Roe's difference splitter for the Euler equations. *Comput. Fluids* 1993; **22**(2/3):215–222.

Deconinck H, Sermeus K and Abgrall R. Status of multidimensional upwind residual distribution schemes and applications in aeronautics. AIAA paper 2000–2328. In *AIAA CFD Conference*, Denver, June 2000.

Degrez G and van der Weide E. Upwind residual distribution schemes for chemical non-equilibrium flows. In *14th AIAA Computational Fluid Dynamics Conference*, Norfolk, June 28 – July 1 1999.

De Palma P, Pascazio G and Napolitano M. An accurate fluctuation splitting scheme for the unsteady two-dimensional Euler equations. *ECCOMAS CFD Conference*, Swansea, September 2001.

De Palma P, Pascazio G, Rossiello G and Napolitano M. A second-order accurate monotone implicit fluctuation splitting scheme for unsteady problems. *J. Comput. Phys.* 2005; **208**(1):1–33.

De Palma P, Pascazio G, Rubino DT and Napolitano M. Residual distribution schemes for advection and advection diffusion problems on quadrilateral cells. *J. Comput. Phys.* 2006; **218**(1):159–199.

Dobes J. *Numerical Algorithms for the Computation of Unsteady Compressible Flow Over Moving Geometries – Application to Fluid-Structure Interaction*. PhD thesis, Université Libre de Bruxelles, 2007.

Dobeš J and Deconinck H. Second order blended multidimensional residual distribution scheme for steady and unsteady computations. *J. Comput. Appl. Math.* 2006; doi:10.1016/j.cam.2006.03.046.

Donea J and Huerta A. *Finite Element Methods for Flow Problems*. John Wiley & Sons, 2003.

Eskilsonn C and Sherwin S. An introduction to spectral/HP finite elements methods for hyperbolic problems. *VKI LS, 34*rd *Computational Fluid Dynamics Course*. von Karman Institute for Fluid Dynamics, 2005.

Evans LC. *Partial Differential Equations*. AMS Press, 1998.

Ferrante A and Deconinck H. *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.*

Godunov SK. An interesting class of quasi-linear systems. *Dokl. Akad. Nauk.* 1961; **139**:521–523.

Henriques JCC and Gato LMC. Use of a residual distribution Euler solver to study the occurrence of transonic flow in wells turbine rotor blades. *Comput. Mech.* 2002; **29**(3):243–253.

Henriques JCC and Gato LMC. A multidimensional upwind matrix distribution scheme for conservative laws. *Comput. Fluids* 2004; **33**:755–769.

Hirsch Ch, Lacor Ch and Deconinck H. *Convection algorithms based on a diagonalization procedure for the multidimensional Euler equations*. AIAA 14th Computational Fluid Dynamics Conference, Norfolk, VA, USA, 667–676, 1987.

Huang LC. Pseudo-unsteady difference schemes for discontinuous solutions of steady-state one dimensional fluid dynamics problems. *J. Comput. Phys.* 1981; **42**:195–211.

Hubbard ME. Non-oscillatory third order fluctuation splitting schemes for steady scalar conservation laws. *J. Comput. Phys.* 2007; **22**(2):740–768.

Hubbard ME and Baines MJ. Conservative multidimensional upwinding for the steady two-dimensional shallow-water equations. *J. Comput. Phys.* 1997; **138**:419–448.

Hubbard ME and Laird AL. High order fluctuation splitting schemes for time-dependent advection on unstructured grids. *Comput. Fluids* 2005; **34**(4/5):443–459.

Hubbard M and Roe PL. Compact high resolution algorithms for time dependent advection problems on unstructured grids. *Int. J. Numer. Methods Fluids* 2000; **33**(5):711–736.

Hughes TJR and Brook A. Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Meth. Appl. Mech. Eng.* 1982; **32**:199–259.

Hughes TJR and Mallet M. A new finite element formulation for CFD III: the generalized streamline operator for multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Eng.* 1986; **58**:305–328.

Johnson C. *Numerical Solution of Partial Differential equations by the Finite Element Method*. Cambridge University Press: Cambridge, 1987.

Kurganov A and Tadmor E. Solution of two-dimensional Riemann problems without Riemann solvers. *Numer. Meth. Part. Diff. Eq.* 2002; **18**:548–608.

Lerat A and Corre C. A residual-based compact scheme for the compressible Navier-Stokes equations. *J. Comput. Phys.* 2001; **170**(2):642–675.

Lerat A and Corre C. Residual-based compact schemes for multidimensional hyperbolic systems of conservation laws. *Comput. Fluids* 2002; **31**(4–7):639–661.

LeVeque RJ. Balancing source terms and flux gradients in high-resolution godunov method: the quasi-steady wave propagation algorithm. *J. Comp. Phys.* 1998; **146**:346–365.

Maerz J and Degrez G. Improving Time Accuracy of Residual Distribution Schemes. Technical Report VKI-PR 96-17, von Karman Institute for Fluid Dynamics, 1996.

Mesaros L. *Multi-Dimensional Fluctuation-Splitting Schemes for the Euler Equations on Unstructured Grids*. PhD thesis, University of Michigan, 1995.

Ni R-H. A multiple grid scheme for solving the Euler equations. *AIAA J.* 1981; **20**:1565–1571.

Nishikawa H. High-order discretization of diffusion terms in residual-distribution methods. *VKI LS, 34*rd *Computational Fluid Dynamics Course*, von Karman Institute for Fluid Dynamics, 2005.

Nishikawa H, Rad M and Roe PL. *A third-order fluctuation splitting scheme that preserves potential flow*. In 15th AIAA Computational Fluid Dynamics Conference, Anaheim, June 2001.

Nishikawa H and Roe PL. On high-order fluctuation splitting schemes for Navier-Stokes equations. *ICCFD3 Proceedings*. Springer-Verlag, 2004.

Nishikawa H and Roe PL. *Adaptive-quadrature fluctuation-splitting schemes for the Euler equations*. AIAA-2005-4865. In 17th AIAA Computational Fluid Dynamics Conference, Toronto, June 2005a.

Nishikawa H and Roe PL. *Towards high-order fluctuation-splitting schemes for Navier-Stokes equations*. AIAA-2005-5244. In 17th AIAA Computational Fluid Dynamics Conference, Toronto, June 2005b.

Nishikawa H and Roe PL. High-order fluctuation splitting schemes for advection-diffusion equations. *ICCFD4 Proceedings*. Springer-Verlag, 2006.

Paillère H. *Multidimensional Upwind Residual Discretization Schemes for the Euler and Navier-Stokes Equations on Unstructured Meshes*. PhD thesis, Université Libre de Bruxelles, 1995.

Paillère H, Corre C and Garcia J. On the extension of the AUSM + scheme to compressible two-fluid models. *Comput. Fluids* 2003; **32**(6):891–916.

Paillère H, Deconinck H and Roe PL. *Conservative upwind residual-distribution schemes based on the steady characteristics of the Euler equations*. AIAA-95-1700. In 12th AIAA Computational Fluid Dynamics Conference, San Diego, 1995.

Paillère H, Degrez G and Deconinck H. Multidimensional upwind schemes for the shallow-water equations. *Int. J. Numer. Methods Fluids* 1998; **26**:987–1000.

Parpia IH and Michalec DJ. Grid-independent upwind scheme for multidimensional flow. *AIAA J.* 1993; **31**(4):646–651.

Powell KG, van Leer B and Roe PL. Towards a genuinely multidimensional upwind scheme. *VKI LS 1990-03, Computational Fluid Dynamics*, von Karman Institute for Fluid Dynamics, 1990.

Quintino T, Ricchiuto M, Csík A and Deconinck H. *Conservative multidimensional upwind residual distribution schemes for arbitrary finite elements*. *ICCFD2 International Conference on Computational Fluid Dynamics 2*. Springer-Verlag, 2002; 88–93.

Ricchiuto M. Contruction and Analysis of Compact Residual Discretizations for Conservation Laws on Unstructured Meshes. PhD thesis, von Karman Institute for Fluid Dynamics and Université Libre de Bruxelles, June ISBN: 2-930389-1. 2005, 6–8.

Ricchiuto M and Abgrall R. Stable and convergent residual distribution for time-dependent conservation laws. *ICCFD4 Proceedings*. Springer-Verlag, 2006.

Ricchiuto M, Abgrall R and Deconinck H. *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.

Ricchiuto M, Abgrall R and Deconinck H. Application of conservative residual distribution schemes to the solution of the shallow water equations on unstructured meshes. *J. Comput. Phys.* 2007; **222**(1):287–331.

Ricchiuto M, Csík A and Deconinck H. Residual distribution for general time dependent conservation laws. *J. Comput. Phys.* 2005; **209**(1):249–289.

Ricchiuto M and Deconinck H. Time-Accurate Solution of Hyperbolic Partial Differential Equations Using Zalesak 1979 and Residual Distribution. Technical Report VKI-SR 99-33, von Karman Institute for Fluid Dynamics, 1999.

Ricchiuto M and Deconinck H. Multidimensional upwinding and source terms in inhomogeneous conservation laws: the scalar case. In *Finite Volumes for Complex Applications III*, Herbin R and Kroner D (eds). HERMES Science Publishing: London, 2002.

Ricchiuto M, Rubino DT, Witteveen JAS and Deconinck H. *A residual distributive approach for one-dimensional two-fluid models and its relation with godunov finite volume schemes*. In *Proceedings of the International Workshop on Advanced Numerical Methods for Multidimensional Simulation of Two-Phase Flow*. Garching, 2003.

Ricchiuto M, Villedieu N, Abgrall R and Deconinck H. High order residual distribution schemes: discontinuity capturing crosswind dissipation and extension to advection diffusion. *VKI LS, 34rd Computational Fluid dynamics Course*. von Karman Institute for Fluid Dynamics, 2005.

Ricchiuto M, Villedieu N, Abgrall R and Deconinck H. On uniformly high-order accurate residual distribution schemes for advection-diffusion. *J. Comput. Appl. Math.* 2007; doi:10.1016/j.cam.2006.03.059.

Roe PL. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.* 1981; **43**:357–372.

Roe PL. Fluctuations and signals – a framework for numerical evolution problems. In *Numerical Methods for Fluids Dynamics*, Morton KW and Baines MJ (eds). Academic Press, 1982; 219–257.

Roe PL. Characteristics based schemes for the Euler equations. *Annu. Rev. Fluid Mech.* 1986a; **18**:337–365.

Roe PL. Discrete models for the numerical analysis of time-dependent multidimensional gas dynamics. *J. Comp. Phys.* 1986b; **63**:458–476.

Roe PL. Linear Advection Schemes on Triangular Meshes. Technical Report CoA 8720, Cranfield Institute of Technology, 1987.

Roe PL and Nishikawa H. Adaptive grid generation by minimising residuals. *Int. J. Numer. Methods Fluids* 2002; **40**:121–136.

Roe PL and Sidilkover D. Optimum positive linear schemes for advection in two and three dimensions. *SIAM J. Numer. Anal.* 1992; **29**(6):1542–1568.

Rossiello G, De Palma P, Pascazio G and Napolitano M. Third-order-accurate fluctuation splitting schemes for unsteady hyperbolic problems. *J. Comput. Phys.* 2007; **222**(1):332–352, doi: 10.1016/j.jcp.2006.07.027.

Rubino DT. *Residual Distribution Schemes for Advection and Advection-Diffusion Problems on Quadrilateral and Hybrid Meshes*. PhD thesis, Politecnico di Bari, 2006.

Rudgyard M. *Cell-Vertex Methods for Steady Inviscid Flow*. VKI LS 1993-04, Computational Fluid Dynamics, von Karman Institute for Fluid Dynamics, 1993.

Rusanov VV. Calculation of interaction of non-steady shock waves with obstacles. *J. Comp. Math. Phys. USSR* 1961; **1**:267–279.

Seaïd M. Non-oscillatory relaxation methods for the shallow-water equations in one and two space dimensions. *Int. J. Numer. Methods Fluids.* 2004; **46**:457–484.

Sermeus K and Deconinck H. An entropy fix for multidimensional upwind residual distribution schemes. *Comput. Fluids* 2005; **34**(4):617–640.

Serre D. *Systems of conservation laws I - Hyperbolicity, Entropies, Shock waves*. Cambridge University Press, Cambridge, UK, 1999.

Sidilkover D and Roe PL. Unification of Some Advection Schemes in Two Dimensions. Technical Report 95-10, ICASE,1995.

Spekreijse SP. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comp.* 1987; **49**:135–155.

Struijs R. *A Multi-Dimensional Upwind Discretization Method for the Euler Equations on Unstructured Grids*. PhD thesis, University of Delft, Netherlands, 1994.

Struijs R, Deconinck H and Roe PL. Fluctuation splitting schemes for the 2D Euler equations. *VKI-LS 1991-01, Computational Fluid Dynamics*, von Karman Institute for Fluid Dynamics, 1991.

Szepessy A. *Convergence of the Streamline Diffusion Finite Element Method for Conservation Laws*. PhD thesis, University of Göteborg, 1989.

Tadmor E. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numer.* 2003; **12**:451–512.

Van Ransbeeck P and Hirsch C. *Multidimensional Upwind Schemes for the Euler/Navier-Stokes Equations on Structured Grids*, Vol. 57 of *Notes on Numerical Fluid Mechanics*. Vieweg Verlag, 1996; 305–338.

van der Weide E and Deconinck H. Positive matrix distribution schemes for hyperbolic systems. *Computational Fluid Dynamics*. John Wiley & Sons: New York, 1996; 747–753.

van der Weide E and Deconinck H. Matrix distribution schemes for the system of Euler equations. In *Euler and Navier-Stokes Solvers Using Multidimensional Upwind Schemes and Multigrid Acceleration*, Vol. 57 of *Notes on Numerical Fluid Dynamics*, Deconinck H and Koren B (eds). Vieweg, 1997; 113–135.

van der Weide E, Deconinck H, Issmann E and Degrez G. A parallel implicit multidimensional upwind residual distribution method for the Navier-Stokes equations on unstructured grids. *Comput. Mech.* 1999; **23**(2):199–208.

Woodward PR and Colella P. The numerical simulation of two-dimensional flows with strong shocks. *J. Comput. Phys.* 1984; **54**:115–173.

Wu L and Bogy DB. Numerical simulation of the slider air bearing problem of hard disk drives by two multidimensional upwind residual distribution schemes over unstructured triangular meshes. *J. Comput. Phys.* 2001; **172**(2):640–657.

Xing Y and Shu C-W. High-order finite difference WENO schemes with the exact conservation property for the shallow-water equations. *J. Comput. Phys.* 2005; **208**(1): 206–227.

Zalesak ST. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* 1979; **31**: 335–362.