Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Computational Physics 214 (2006) 773–808
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Essentially non-oscillatory Residual Distribution schemes for hyperbolic problems

R. Abgrall *

Institut Universitaire de France and Mathématiques Appliquées de Bordeaux, Projet Scalapplix, INRIA FutURs, Université Bordeaux I, 351 cours de la Libération, F-33405 Talence Cedex, France

Received 3 June 2005; received in revised form 22 September 2005; accepted 14 October 2005
Available online 3 January 2006

Abstract

The Residual Distribution (RD) schemes are an alternative to standard high order accurate finite volume schemes. They have several advantages: a better accuracy, a much more compact stencil, easy parallelization. However, they face several problems, at least for steady problems which are the only cases considered here. The solution is obtained via an iterative method. The iterative convergence must be good in order to get spatially accurate solutions, as suggested by the few theoretical results available for the RD schemes. In many cases, especially for systems, the iterative convergence is not sufficient to guaranty the theoretical accuracy. In fact, up to our knowledge, the iterative convergence is correct in only two cases: for first order monotone schemes and the (scalar) Struij's PSI scheme which is a multidimensional upwind scheme. Up to our knowledge, the iterative convergence is poor for systems, except for the blended scheme of Deconinck et al. [Á. Csík, M. Ricchiuto, H. Deconinck, A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws, J. Comput. Phys. 179(2) (2002) 286–312] and Abgrall [R. Abgrall, Toward the ultimate conservative scheme: following the quest, J. Comput. Phys. 167(2) (2001) 277–315] which are also a genuinely multidimensional upwind scheme.

A second drawback is that their construction relies, up to now, on a single first order scheme: the N scheme. However, it is known that standard first order finite volume schemes can be rephrased into a Residual Distribution framework. Unfortunately, the standard way of upgrading the order of accuracy to second order leads to very unsatisfactory results but clearly the construction of good schemes based on a wider class of first order schemes would be interesting.

In this paper, we analyze these two problems, and show they are linked. We propose a fix and demonstrate its efficiency on several test cases that cover a wide range of applications. Our solution extends considerably the number of working RD schemes.

1. Introduction

We are interested in the numerical solution of hyperbolic problems (scalar or system) on unstructured meshes, with a particular emphasis on the Euler equations for fluid mechanics. Many type of methods are

* Tel.: +33 5 56 84 60 68; fax: +33 5 56 84 26 26.
E-mail address: abgrall@math.u-bordeaux.fr.

0021-9991/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2005.10.034

available in the literature, such as the high resolution finite volume schemes or the stabilized finite element methods. Here, we are interested in a particular class of schemes that share similarities with the high resolution finite volume schemes and the stabilized FEM: they possess a nonlinear and non-oscillatory mechanism that is inspired in part by what is done for high resolution finite volume schemes. They can also be interpreted as continuous finite element methods in which the test functions depend on the solution. Hence a variational formulation exist, and thanks to this interpretation, they share the residual property: the exact solution, if smooth enough, satisfies this variational formulation. Thanks to this, they are very accurate. Their implementation can also be done in a *very* compact way, hence the parallelization is simple.

The non-oscillatory property of the RD schemes are obtained in two steps. First, a low order monotone non-oscillatory scheme is written. Second, a high order scheme is constructed from it by enforcing the non-oscillatory property by comparison with the low order scheme. This is done thanks to a trick that uses in deep the structure of PDE. We review in some details the construction of RD schemes in Section 2. High order extensions also exist, see [2] as well as extensions for unsteady problems [3,4].

However, these schemes suffer two major drawbacks. One of them is that, for steady problems, the numerical solution is obtained via an iterative scheme that, in general, does not converge. More precisely, the iterative procedure has a nice convergence behavior in only two cases: the case of dissipative first order schemes such as the N scheme and a Lax–Friedrichs type scheme that we recall later in the text, and the case of multidimensional upwind schemes such as the N scheme. In the case of systems and up to our knowledge, the iterative behavior is good in the case of blended schemes such as those described in [5,1]; but these schemes are not robust enough to serve as all-purpose solvers. In the case of the second order schemes of [3], as mentioned in this reference, the iterative convergence is bad. If one looks at the spatial structure of the local residuals, relatively high values are obtained at apparently randomly distributed locations. In the unsteady case, the solution at each time step is also obtained by an iterative method that has no good converge properties. In these two cases, the formal theoretical accuracy of the RD scheme becomes problematic, because the accuracy is guaranteed *at convergence* only or if the spurious residual is small enough. However, the results are good in practice.

Another drawback is the relative lack of flexibility of the technique. Up to now, the *only* first order scheme that produces successful high order schemes is the so-called N scheme [6]. Since it is easy to see that any finite volume scheme can be rephrased as a RD scheme, it is very tempting to consider any first order finite volume scheme and apply the construction on it. This would provide an elegant way of ''exporting'' the properties of the FV schemes to this setting such as the positivity preserving properties of some of these schemes. However, the result is very disappointing, as we see in Section 3!

The purpose of this paper is to analyze these phenomena that seem to be linked, and to propose effective solutions. This analysis is carried out in Section 4. In order to solve the two problems (iterative convergence and flexibility of the method), it seems that we need to lose the rigorous non-oscillatory property of the RD schemes. However, even if the schemes are not anymore strictly non-oscillatory, the schemes seem to work extremely well in practice. We demonstrate this on several examples for scalar and system cases (Euler equations of fluid mechanics).

The paper is organized as follows. First, we introduce the Residual Distribution schemes, provide the design principles and several examples. In the second section, we illustrate by several experiments the problems of the second order schemes: lack of convergence, wiggly behavior. Part 3 is an attempt to analyze these difficulties, and provide a fix. The last section is devoted to the intensive evaluation of our solution. A conclusion ends the text.

2. The Residual Distribution formalism

2.1. General considerations

We consider the following hyperbolic problem:

$$\begin{aligned} &\operatorname{div}\mathbf{f}(\mathbf{u}) = 0, \quad x \in \Omega, \\ &\mathbf{u} = g \text{ weakly}, \quad x \in \partial\Omega, \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $\mathbf{u} : \Omega \to \mathbb{R}^m$ and $\mathbf{f}$ is a regular function defined on an open subset of $\mathbb{R}^m$. The function g is regular enough for the boundary condition to have a meaning. The set $\partial\Omega^-$ is the inflow part of $\partial\Omega$.

For the sake of simplicity, we assume that Ω is polygonal, and we consider $\mathcal{T}_h$ a shape regular conformal triangulation of Ω. For the sake of simplicity, we also assume $d = 2$, the discussion can easily be adapted to $d = 3$.

We denote by $\mathcal{V} = \{M_j\}_{j=1,\dots,n_s}$ the vertices of $\mathcal{T}_h$, and $\{T_j\}_{j=1,\dots,n_t}$ the set of triangles of $\mathcal{T}_h$. The vertices of T are M_{j_1}, M_{j_2} and M_{j_3}. Most of the time, we denote them by their index in the list $\mathcal{V}$, and when there is no ambiguity, they are simply denoted by 1, 2 and 3. Last, $\mathcal{V}(i)$ denotes the set of vertices that are connected to M_i by one edge of $\mathcal{T}_h$. The parameter h denotes the supremum over the triangles of $\mathcal{T}_h$ of $\sqrt{|T|}$.

In the RD schemes, the solution of (1) is approximated at the vertices: the numerical approximation is represented by $(\mathbf{u}_j)_{j=1,\dots,n_s}$. From this we construct a continuous interpolant $\mathbf{u}^h$: the function is linear on each triangle T and $\mathbf{u}^h(M_j) = \mathbf{u}_j$.

In each triangle T, we assume to have in hand residuals $(\Phi_j^T)_{M_j \text{ vertex of } T}$, $\Phi_j^T := \Phi_j^T(\mathbf{u}^h)$, such that the following conservation relation holds:

$$\sum_{j \in T} \Phi_j^T = \int_{\partial T} \mathbf{f}(\mathbf{u}^h(x))\, \mathrm{d}\partial T := \Phi^T. \tag{2}$$

The quantity Φ^T is called the total residual over T. We show latter several examples of such decompositions.

Once this is done, the RD scheme writes

$$\text{for all } M_j, \qquad \sum_{T \text{ such that } M_j \in T} \Phi_j^T(\mathbf{u}^h) = 0. \tag{3}$$

In order to simplify the text, we skip the general problem of setting up boundary condition. This point is addressed later in Section 2.4. Similarly, (3) is rewritten as $\sum_{T \ni M_j}$ or $\sum_{T \ni j}$

The relation (3) raises three questions:

1. How to define the residuals Φ_j^T?
2. Which design principles should be applied, in particular how to get accurate and non-oscillatory results?
3. How to solve (3)?

2.2. Design principles

Until the end of the paper, we drop the superscript T in the writing of residuals when there is no ambiguity, and when it is unnecessary, thanks to the form (3) of the RD schemes.

2.2.1. Consistency with (1)

Consider a sequence of shape regular meshes $\{\mathcal{T}_h\}$ with $h \to 0$. We assume that we can solve (3) exactly, the solution is denoted by $\mathbf{u}^h = \{\mathbf{u}_j^h\}_{j=1,\dots,n_s}$. In [7], we show that provided the solution satisfies

1. there exists a constant C such that for all h, $\max_j \|\mathbf{u}_j^h\| \leqslant C$;
2. there exists a function $\mathbf{u}$ in $L^2_{\mathrm{loc}}(\mathbb{R}^m)^d$ and a subsequence of $\{\mathbf{u}^h\}$ such that $\mathbf{u}^h \to \mathbf{u}$ in L^2_{loc};
3. the conservation relation (2) holds true for any $\mathbf{u}^h$,

then $\mathbf{u}$ is a weak solution of (1).

This is the analog of the classical Lax–Wendroff theorem. A slightly more general result is shown in [8].

2.2.2. Solution scheme for (3)

Eq. (3) is generally solved by an iterative scheme. In most of the present paper (i.e. except for system cases), we consider the simplest one, namely

$$\text{for all } j, \quad \begin{cases} \mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \omega_j \left(\sum_{T \text{ such that } M_j \in T} \Phi_j^T(\mathbf{u}^n) \right) \\ \mathbf{u}_j^0 \text{ given.} \end{cases} \tag{4}$$

The parameters ω_j are chosen so that the stability of the scheme is ensured. If (4) converges, this defines a solution of (3). The remaining question is its uniqueness. This is a very difficult problem that is at the core of the present paper.

2.2.3. Monotonicity preserving schemes

The monotonicity preserving nature of a scheme can be formalized in the case of a scalar problem (1) and is more intuition-based in the system case.

In the case of a scalar problem, if one assumes that the residuals have the form

$$\Phi_j^T = \sum_{j \in T} c_{ij}^T (\mathbf{u}_i - \mathbf{u}_j), \tag{5}$$

then the scheme (3) writes

$$a_{ii}\mathbf{u}_j = \sum_{j \in \mathcal{V}(i)} a_{ij}\mathbf{u}_j \tag{6}$$

with

$$\begin{aligned} a_{ii} &= \sum_{T, T \ni M_j} \sum_{j \in T} c_{ij}^T, \\ a_{ij} &= \sum_{T, M_i \in T \text{ and } M_j \in T} c_{ij}^T. \end{aligned} \tag{7}$$

All the known examples of RD scheme write in the form (5). In the case of a system problem (1), the c_{ij}^T are matrices and (6) and (7) still hold.

If the coefficients c_{ij}^T are all positive, it is clear that (6) defines a scheme with a maximum principle *provided* there exists a solution. Note that a necessary condition for the existence of a solution is $a_{ii} > 0$ for all $i = 1, \ldots, n_s$. This condition is translated for (4) by

$$a_{ii}\omega_i \leqslant 1.$$

A local condition is

$$\omega_i \leqslant \left\{ \max_{T, T \ni i} \left[\sum_{j \in T} c_{ij}^T \right] \right\}^{-1}.$$

This is the one that is used in practice.

2.2.4. Accuracy: the linearity preserving (LP) condition

Under which condition can the solution of (2) be a second order accurate solution of (1)? We briefly recall the analysis of [5]. It is shown that a *converged* RD scheme (3) produces a formally second order accurate solution of the *steady problem* (1) under the following three requirements:

1. the mesh is regular;
2. the approximation $\mathbf{u}^h$ is second order accurate on smooth solutions;
3. for any smooth solution of (1), $\Phi_i^T = \mathcal{O}(h^3)$ for any vertex M_i and any triangle T such that $M_i \in T$.

For this reason, it is *essential* that Eq. (3) is exact or approximately exact with an error at most $\mathcal{O}(h^3)$ otherwise accuracy is lost.

In most cases, the third condition is met by imposing that there exists a family of uniformly bounded coefficients (or matrices for system problems) $\boldsymbol{\beta}_i^T$ such that

$$\Phi_i^T = \boldsymbol{\beta}_i^T \Phi^T. \tag{8}$$

Indeed, it is easy to show that

$$\int_{\partial T} \mathbf{f}(\mathbf{u}^h)\cdot\vec{n}\,\mathrm{d}\partial T = \mathcal{O}(h^3)$$

when $\mathbf{f}(\mathbf{u}^h)$ is a second order approximation corresponding to a smooth solution and $\vec{n}$ is the outward unit vector of ∂T. The condition (8) is the Linearity Preservation (LP) condition introduced in [6].

It is known that it is not possible to have a linear scheme that is both monotonicity preserving and linearity preserving: this is Godunov theorem [9]. The schemes that satisfy both requirements must be nonlinear. The construction of such schemes is the topic of the next sub-section.

2.2.5. Systematic construction of second order LP schemes

The problem is the following. Considering a triangle T, assume we are given residuals that define a first order[1] monotone scheme, (Φ_1, Φ_2, Φ_3). We want to construct a second order scheme defined by its residuals $(\Phi_1^\star, \Phi_2^\star, \Phi_3^\star)$ such that the resulting scheme is

1. conservative

$$\sum_{i=1}^3 \Phi_i = \sum_{i=1}^3 \Phi_i^\star = \Phi,$$

2. monotonicity preserving,
3. linearity preserving.

We first focus on scalar problems, then sketch a method for systems.

The first remark is that if one defines $x_i = \Phi_i/\Phi$, we notice that

$$\sum_{i=1}^3 x_i = 1.$$

Then we define $\beta_i = \Phi_i^\star/\Phi$, the problem can be reformulated as finding a mapping $(x_1, x_2, x_3) \mapsto (\beta_1, \beta_2, \beta_3)$ such that the scheme is

1. conservative: $\sum_{i=1}^3 \beta_i = 1$.
2. monotonicity preserving: for all $i = 1, 2, 3$, $x_i\beta_i \geqslant 0$. This condition comes from the fact that

$$\Phi_i^\star = \frac{\Phi_i^\star}{\Phi}\frac{\Phi}{\Phi_i}\Phi_i = \frac{\beta_i}{x_i}\sum_{j\neq i} c_{ij}(u_i - u_j) = \sum_{j\neq i} c_{ij}^\star (u_i - u_j)$$

with $c_{ij}^\star = \frac{\beta_i}{x_i} c_{ij}$. Since $c_{ij} \geqslant 0$, the positivity of $c_{ij}^\star$ is equivalent to $x_i\beta_i \geqslant 0$.
3. linearity preserving: we want β_i bounded for any i.

In [2], we provide a geometrical interpretation of these conditions, and several solutions to this problem. We repeat the argument. The key remark is that since $\sum_j x_j = \sum_j \beta_j = 1$, we can interpret the coordinates (x_1, x_2, x_3) and $(\beta_1, \beta_2, \beta_3)$ as the barycentric coordinates of points L and H with respect to an abstract reference triangle (A_1, A_2, A_3) that we choose to be equilateral for symmetry. The points L and H are defined by

$$L = x_1A_1 + x_2A_2 + x_3A_3 \quad \text{or equivalently} \quad \overrightarrow{A_1L} = x_2\overrightarrow{A_1A_2} + x_3\overrightarrow{A_1A_3}$$
$$H = \beta_1A_1 + \beta_2A_2 + \beta_3A_3 \quad \text{or equivalently} \quad \overrightarrow{A_1H} = \beta_2\overrightarrow{A_1A_2} + \beta_3\overrightarrow{A_1A_3}$$

[1] That is, for which we only have $\Phi_i = \mathcal{O}(h^2)$.

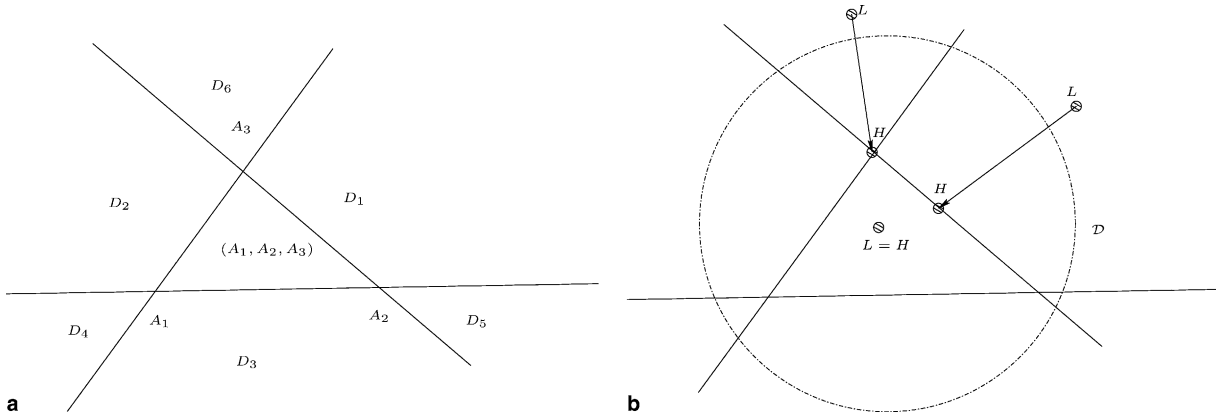

Fig. 1. Geometrical representation of the mapping $(x_1, x_2, x_3) \mapsto (\beta_1, \beta_2, \beta_3)$.

In Fig. 1(a), we have defined seven sub-domains: the triangle (A_1, A_2, A_3) and the six domains D_i. The problem is to find a mapping that project the point L onto a bounded subdomain so that L and H belongs to the same sub-domain. A geometrical representation of a possible projection is given in Fig. 1(b). Note that here, the projection leaves invariant the triangle (A_1, A_2, A_3): we project onto this triangle. What is important is that the coefficients β_j be bounded, so any bounded region can play the role of invariant region onto which the projection is carried out, for example the disk $\mathcal{D}$ of Fig. 1(b). In the next examples, the invariant region is the triangle (A_1, A_2, A_3).

One of these possible projections is the PSI "limiter" first introduced by R. Struijs in his PhD thesis [10], in a different form:

$$\beta_i = \frac{x_i^+}{\sum_j x_j^+}, \tag{9}$$

so that

$$\Phi_i^\star = \beta_i \Phi. \tag{10}$$

We note that there is no difficulty in the definition of β_i (except the fact that Φ may vanish, in which case we set $\Phi_i^\star = 0$) because

$$\sum_j x_j^+ = \sum_j x_j - \sum_j x_j^- \geqslant \sum_j x_j = 1.$$

This construction can be applied to *any* monotone scheme. However, to be valid, one needs to be able to solve the problem (3). A necessary condition is that the coefficient a_{ii} associated to the coefficients $c_{ij}^\star$ by (7) be >0. We come back later to this key point.

We can extend this construction in the system case. This has been done in [11]. We start from (1), and assume to have in hand a first order non-oscillatory scheme. Examples are given in the next section. If $(\mathbf{r}_j)_{j=1,\ldots,d}$ is a basis of $\mathbb{R}^d$, we can decompose the residuals Φ_i as

$$\Phi_i = \sum_{j=1}^{d} \varphi_i^j \mathbf{r}_j. \tag{11}$$

The total residual also admits such a decomposition,

$$\Phi = \sum_{j=1}^{d} \varphi^j \mathbf{r}_j. \tag{12}$$

From the conservation relation (2), we have, for any $j = 1, \ldots, d$,

$$\sum_{i=1}^{m} \varphi_i^j = \varphi^j. \tag{13}$$

Thus we can apply the scalar construction to each set of scalar residuals $\{\varphi_i^j\}_{i=1,m}$ for $j = 1, \ldots, d$. We denote by $(\varphi_i^j)^\star$ the result of the construction.

This enable to define uniformly bounded matrices $\mathbf{B}_i$ such that the LP residuals are

$$\Phi_i^\star = \sum_{j=1}^{m} (\varphi_i^j)^\star \mathbf{r}_j := \mathbf{B}_i \Phi. \tag{14}$$

This scheme, with characteristic variables φ_i^j has been studied in [11] and shown non-oscillatory. The choice of $(\mathbf{r}_j)_{j=1,\ldots,d}$ is discussed in the next section.

2.3. Examples

Many schemes are in fact Residual Distribution schemes. Among the most known, we mention the streamline diffusion method of Johnson and coworkers [12,13], the streamline upwind Petrov–Galerkin (SUPG) and Galerkin least-squares finite element methods of Hughes and coworkers [14,15] and the cell vertex finite volume methods of Ni [16] and Morton et al. [17,18]. Here, we are interested in the construction of oscillation free schemes, we only describe in detail some first order RD schemes.

2.3.1. Some genuinely multidimensional schemes

2.3.1.1. A genuinely multidimensional upwind scheme. To begin with, we consider the scalar problem (1) with a linear flux,

$$\boldsymbol{\lambda} \cdot \nabla u = 0 \tag{15}$$

with inflow boundary conditions. For a piecewise linear interpolation of u, we get

$$\Phi = \int_T \boldsymbol{\lambda} \cdot \nabla u \, \mathrm{d}\mathbf{x} = \sum_{j=1}^{3} k_j u_j,$$

where, if $\vec{n}_j$ represents the scaled inward normal to T opposite to the vertex M_j (see Fig. 2), $k_j = \frac{1}{2} \boldsymbol{\lambda} \cdot \vec{n}_j$.

Roe's N scheme is then defined as

$$\Phi_i = k_i^+ (u_i - \tilde{u}), \tag{16a}$$

where $k_i^+ = \max(k_i, 0)$ and $\tilde{u}$ is defined so that the conservation property holds. A simple algebra shows that

$$\tilde{u} := n \left(\Phi - \sum_j k_j^+ u_j \right) = n \sum_j k_j^- u_j \tag{16b}$$

with $n = (\sum_j k_j^-)^{-1}$ and $k_j^- = \min(k_j, 0)$. Since $\sum_{j=1}^{3} k_j = 0$, there are two possible cases:

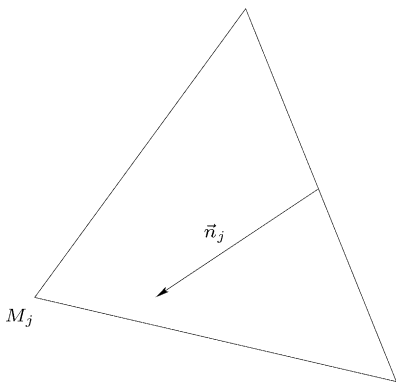

Fig. 2. Illustration of $\vec{n}_j$.

- The one target case: only one k_j is positive, say k_1. We get

$$\Phi_1 = \Phi, \quad \Phi_2 = \Phi_3 = 0.$$

- The two target case: only one k_j is negative, say k_3. In that case, we have $\tilde{u} = u_3$ and

$$\Phi_1 = k_1(u_1 - u_3), \quad \Phi_2 = k_2(u_2 - u_3), \quad \Phi_3 = 0.$$

This scheme is upwind: if $k_j \leqslant 0$ then $\Phi_j = 0$. It has an important property: for any interior vertex M_i, one (and only one) of the triangles surrounding M_i is upwind. This is also true for any vertex of the outflow boundary. Thanks to this, the coefficient a_{ii} of (7) is >0, and (3) leads to a linear system that always has a unique solution. One way of seeing this is there exist a numbering of the nodes by level sets such that the linear system is almost lower triangular. A more rigorous way of seeing that is that the N scheme satisfies an energy inequality, see [19].

In the case of the true nonlinear problem (1), the previous construction can be extended provided a suitable averaged speed $\overline{\boldsymbol{\lambda}}$ can be defined. The N scheme writes as in (16) with $\boldsymbol{\lambda}$ replaced by $\overline{\boldsymbol{\lambda}}$. Thanks to the results recalled in Section 2.2.1, we get easily one constraint on $\overline{\boldsymbol{\lambda}}$, namely that

$$\int_{\partial T} \mathbf{f}(u^h)\,\mathrm{d}\partial T = \sum_{j=1}^{3} \Phi_j = \int_{\partial T} \overline{\boldsymbol{\lambda}} \cdot \vec{n} u^h\,\mathrm{d}\partial T.$$

which is nothing more than the extension of Roe's linearization to the problem (1). Since the interpolant u^h is linear, we have the equivalent characterization

$$\overline{\boldsymbol{\lambda}} := \frac{\int_T \nabla_{\mathbf{u}} \mathbf{f}(\mathbf{u}^h)\,\mathrm{d}\mathbf{x}}{|T|}. \tag{17}$$

Note that the numbering of the level sets, as in the constant velocity case, can be done similarly as long as $\|\overline{\boldsymbol{\lambda}}\| > 0$. The second order extension of the N scheme is Struijs' PSI scheme [10] that we denote by N-PSI in this paper. It uses (9).

The N scheme has a system version, that was introduced by [20] and analyzed in [19]. In the case of an hyperbolic linear problem

$$\mathbf{A}\frac{\partial \mathbf{u}}{\partial x} + \mathbf{B}\frac{\partial \mathbf{u}}{\partial y} = 0, \tag{18}$$

given any direction $\vec{n}$, we can define the positive and negative parts of the matrix $\mathbf{K}_{\vec{n}}$, defined by

$$\mathbf{K}_{\vec{n}} := \mathbf{A}n_x + \mathbf{B}n_y.$$

This matrix is also sometimes denoted as $\mathbf{K}_{\vec{n}} = (\mathbf{A}, \mathbf{B}) \cdot \vec{n}$. Using the fact that the three scaled inward normals to T, $\vec{n}_1$, $\vec{n}_2$, $\vec{n}_3$, sum up to 0, we can define the system N scheme as

$$\Phi_i = \mathbf{K}_i^+(\mathbf{u}_i - \tilde{\mathbf{u}}) \tag{19a}$$

with

$$\tilde{\mathbf{u}} := \mathbf{N}\left(\Phi - \sum_{j=1}^{3} \mathbf{K}_j^+ \mathbf{u}_j\right) = \mathbf{N}\left(\sum_{j=1}^{3} \mathbf{K}_j^- \mathbf{u}_j\right) \tag{19b}$$

and

$$\mathbf{N} := \left(\sum_{j=1}^{3} \mathbf{K}_j^-\right)^{-1}. \tag{19c}$$

Here we have simplified the notation $\mathbf{K}_{\vec{n}_j}$ into $\mathbf{K}_j$. In [19], we show that if (18) is symetrizable, the matrices $\mathbf{N}\mathbf{K}_j^-$ can be defined even if $\sum_{j=1}^3 \mathbf{K}_j^-$ is not invertible.

The scheme can be generalized to nonlinear problems by a simple extension of (17), see for example [21] in the case of the Euler equation with the equation of state of a perfect gaz and γ constant. The difficult question is to know whether or not the linearized system is hyperbolic. In the case of the Euler equation and Roe–Struijs–Deconinck linearization, this is true, but no answer can be given in the general case. See however [19] for a different approach, and [1] for a very interesting approximate linearization leading to a conservative system.

The high order extension is carried out as in Section 2.2.5. The basis used in (11)–(14) are in practical applications of two types. Either we chose a direction $\vec{n}$, say the velocity direction, and define the basis as the eigenvectors of $\mathbf{K}_{\vec{n}}$. Or we simply choose the canonical basis of $\mathbb{R}^m$, i.e. we proceed the high order construction component by component. We refer to [3] for the discussion. Starting from the system N scheme, with the characteristic decomposition based on the velocity and the PSI limiter (9), we get the so-called PSI system N scheme, see [3], still denoted by N-PSI in this paper.

2.3.1.2. A Lax–Friedrich type scheme. The one dimensional version of the Lax–Friedrich scheme, for the one dimensional version of (1) writes

$$\begin{aligned} &\hat{f}_{i+1/2} - \hat{f}_{i-1/2} = 0, \\ &u = g \text{ on the inflow boundary} \end{aligned} \tag{20}$$

with

$$\hat{f}_{i+1/2} = \frac{1}{2}\left(f(u_{i+1}) + f(u_i) - \alpha_{i+1/2}(u_{i+1} - u_i)\right)$$

and $\alpha_{i+1/2} \geqslant 0$ suitably chosen.

The first relation of (20) can be rewritten as

$$\frac{1}{2}\left(f(u_{i+1}) - f(u_i) - \alpha_{i+1/2}(u_{i+1} - u_i)\right) + \frac{1}{2}\left(f(u_i) - f(u_{i-1}) + \alpha_{i-1/2}(u_i - u_{i-1})\right) = 0,$$

that is

$$\phi_i^{i+1/2} + \phi_i^{i-1/2} = 0$$

with, for any j,

$$\begin{aligned} \phi_j^{j+1/2} &= \frac{1}{2}\left(f(u_{j+1}) - f(u_j) + \alpha_{i+1/2}(u_j - u_{j+1})\right), \\ \phi_{j+1}^{j+1/2} &= \frac{1}{2}\left(f(u_{j+1}) - f(u_j) + \alpha_{j+1/2}(u_{j+1} - u_j)\right). \end{aligned}$$

The natural two dimensional generalization of this is

$$\Phi_j^T = \frac{1}{3}\left(\Phi^T + \alpha_T\left[\sum_{k\in T}(u_j - u_k)\right]\right). \tag{21}$$

Clearly, the conservation property (2) holds. The scalar version is monotone provided

$$\alpha_T \geqslant \max_{j=1,3} |k_j|$$

for scalar problems and $\alpha_T = \max_{j=1,3}\rho(K_j)$ for systems. Here $\rho(A)$ denotes the spectral radius of the matrix A, the matrices K_i are evaluated from the Jacobian matrices of the Euler flux evaluated at the Roe average [21]. In this case, γ, the ratio of specific heats, is constant.

Last, when $f(u) = \boldsymbol{\lambda} u$, the scheme has the following local energy structure:

$$\sum_{j\in T} u_j \Phi_j^T = \frac{1}{2}\int_{\partial T} \boldsymbol{\lambda}\cdot\mathbf{n} u \, \mathrm{d}\partial T + \frac{\alpha_T}{2}\sum_{i,j\in T}(u_i - u_j)^2.$$

This can easily be extended to more general scalar fluxes as well to symetrizable hyperbolic systems.

2.3.2. Some non-genuinely multidimensional schemes

In contrast to the previous examples where the directions needed to construct the residual could not be associated to the geometry of some control volume, here, we consider examples where the construction is done by considering a control volume and the associated directions. Two set of examples are described. In the first one, that we denote by some abuse of language "finite volume schemes", the construction starts from a standard one dimensional flux. In the second example, a generalization of Roe's one dimensional scheme is considered.

2.3.2.1. Finite volume schemes. Any finite volume type scheme can be rephrased as a RD scheme. The interest of this remark is to considerably extend, in theory, the number of RD schemes, and in particular new high order schemes can be constructed using the technique of Section 2.2.5. For example, starting from a positivity preserving scheme, and doing the high order extension component by component, it becomes possible to construct, for the Euler equations in fluid mechanics, a LP density-positivity preserving scheme. This may be interesting because it is not clear at all that the system N-PSI is density-positivity preserving, even though this scheme has been shown very robust experimentally.

For any vertex M_i of $\mathscr{T}_h$, we consider the dual control volume $\mathscr{C}_i$ which is constructed by connecting, for each triangle surrounding M_i, its centroid and the mid-points of the edges containing M_i, see Fig. 3.

Now, let us consider a consistent flux $\mathscr{F}$. The finite volume approximation of (1) writes

$$\sum_{j\in\mathscr{V}(i)}\left(\mathscr{F}(\mathbf{u}_i,\mathbf{u}_j,\vec{n}_{ij}^{T_{\rm up}})+\mathscr{F}(\mathbf{u}_i,\mathbf{u}_j,\vec{n}_{ij}^{T_{\rm down}})\right)=0, \tag{22}$$

where the triangles $T_{\rm up}$ and $T_{\rm down}$, for the edge $[i,j]$, are defined in Fig. 3. Here we consider a first order scheme for the sake of simplicity, but also because it is the only interesting case for our purpose in this paper.

Instead of summing up over the edges in (22), we can sum up over the triangles around M_i,

$$\sum_{T,M_i\in T}\left(\mathscr{F}(\mathbf{u}_i,\mathbf{u}_j,\vec{n}_{ij}^{T})+\mathscr{F}(\mathbf{u}_i,\mathbf{u}_k,\vec{n}_{\rm ik}^{T})\right)=0 \tag{23}$$

with the notations of Fig. 3. Since the boundary of $\mathscr{C}_i$ is closed,

$$\sum_{T,M_i\in T}\left(\mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{ij}^{T}+\mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{\rm ik}^{T}\right)=0,$$

we have

$$\sum_{T,M_i\in T}\left(\mathscr{F}(\mathbf{u}_i,\mathbf{u}_j,\vec{n}_{ij}^{T})+\mathscr{F}(\mathbf{u}_i,\mathbf{u}_k,\vec{n}_{\rm ik}^{T})-\mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{ij}^{T}-\mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{\rm ik}^{T}\right)=0. \tag{24}$$

We set

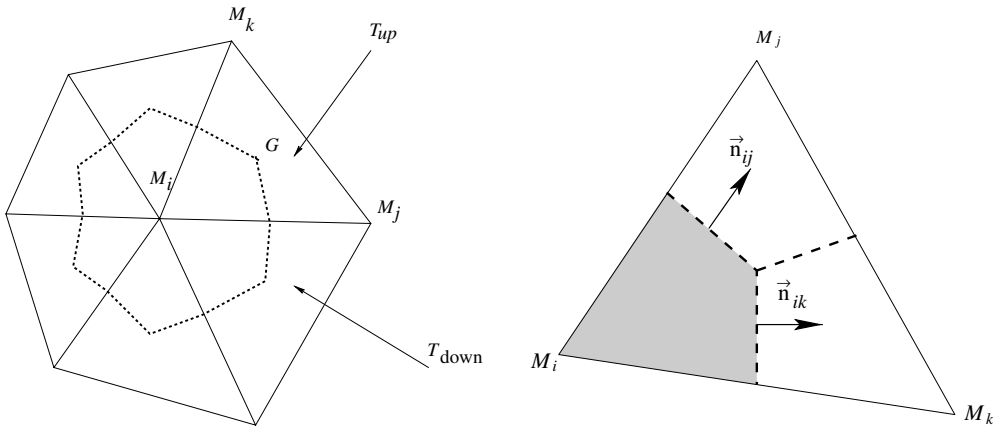

Fig. 3. Geometrical elements for the dual cell of M_i.

$$\Phi_i^T := \mathscr{F}(\mathbf{u}_i, \mathbf{u}_j, \vec{n}_{ij}^T) + \mathscr{F}(\mathbf{u}_i, \mathbf{u}_k, \vec{n}_{\mathrm{ik}}^T) - \mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{ij} - \mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_{\mathrm{ik}}$$
$$= \mathscr{F}(\mathbf{u}_i, \mathbf{u}_j, \vec{n}_{ij}^T) + \mathscr{F}(\mathbf{u}_i, \mathbf{u}_k, \vec{n}_{\mathrm{ik}}^T) + \mathbf{f}(\mathbf{u}_i)\cdot\frac{\vec{n}_i}{2} \tag{25}$$

and because of the construction of $\mathscr{C}_i$, we see that

$$\sum_{M_j\in T}\Phi_j^T = \frac{1}{2}\sum_{M_j\in T}\mathbf{f}(\mathbf{u}_i)\cdot\vec{n}_i. \tag{26}$$

Here, we slightly extend the definition (2) by

$$\sum_{M_j\in T}\Phi_j^T = \int_{\partial T}(\mathbf{f}(\mathbf{u}))^h\cdot\vec{n}_i\,\mathrm{d}\partial T = \Phi^T, \tag{27}$$

i.e. make a piecewise linear interpolation of the flux $\mathbf{f}(\mathbf{u})$. The results of Section 2.2.1 can be extended to this case, see [5].

In this paper, we use the finite volume scheme with Roe's flux [22].

2.3.2.2. A multidimensional version of Roe' scheme for the Euler equations. This version has been first presented in [20]. Using once more the Roe average of [21], we do the same construction as in the finite volume case with

$$\mathscr{F}(U_1, U_2, \vec{n}) = ((\overline{\mathbf{A}}, \overline{\mathbf{B}})\cdot\vec{n})^+ U_1 + ((\overline{\mathbf{A}}, \overline{\mathbf{B}})\cdot\vec{n})^- U_2. \tag{28}$$

The conservation property is not guarantied by edge but on the triangle since the sum of the residuals constructed from (28) is

$$\int_{\partial T}\mathbf{f}(\mathbf{u}^h)\cdot\vec{n}\,\mathrm{d}\partial T,$$

where $\mathbf{u}^h$ is obtained by interpolating the Roe parameter vector

$$Z = \sqrt{\rho}(1, \vec{u}, H)^T$$

which is linearly interpolated in T. An exact linearization is obtained, as in the one dimensional case, because $\mathbf{u}$ as well as $\mathbf{f}$ are quadratic in Z. Thus, in (28), $(\overline{\mathbf{A}}, \overline{\mathbf{B}})$ are the Jacobian matrices evaluated at the average state defined by $\overline{Z} = (Z_1 + Z_2 + Z_3)/3$. See [20] for more details.

2.4. Treatment of inflow boundary conditions

We consider the simplified problem

$$\begin{aligned} &\boldsymbol{\lambda}\cdot\nabla u = 0 \quad \text{in } \Omega,\\ &u = g \qquad\quad \text{on } \Gamma^-. \end{aligned} \tag{29}$$

Assuming that g is the restriction of a sufficiently regular function, still denoted by g, and defined on Ω, it is known that (29) has a unique solution.

We consider the linear preserving scheme of Section 2.2.5: we have

$$\Phi_i^T = \beta_i^T\int_T \boldsymbol{\lambda}\cdot\nabla u\,\mathrm{d}\mathbf{x}$$

with β_i^T uniformly bounded. If φ_i is the piecewise linear hat function for which $\varphi_i(M_j) = \delta_i^j$, we can write

$$\Phi_i^T = \int_T \omega_i\boldsymbol{\lambda}\cdot\nabla u\,\mathrm{d}\mathbf{x}$$

with ω_i defined on any triangle by $(\omega_i)_{|T} = \varphi_i + (\beta_i^T - \frac{1}{3})$. We introduce the spaces

$$V_h = \{v \text{ continuous, } v \text{ linear on each triangle, } v_{\Gamma^-} = g\},$$
$$W_h = \text{span}(\omega_1, \ldots, \omega_{ns}).$$

The scheme (3) can be rewritten in an abstract form: find $u^h \in V_h$ such that for any $w^h \in W_h$,

$$\int_\Omega w^h \boldsymbol{\lambda} \cdot \nabla u^h \, d\mathbf{x} = 0. \tag{30}$$

Using (30), the interpretation of the boundary conditions becomes clear: If M is any vertex that belongs to a triangle which intersect Γ^-, for example the points i or j in Fig. 4, we have

$$\sum_{T \ni M} \int_T \omega_M \boldsymbol{\lambda} \cdot \nabla u^h \, d\mathbf{x} = 0.$$

If $T \ni M$ and T intersects Γ^- (examples are the triangles T_1 and T_2 of Fig. 4), we have the three equivalent formulations

$$\begin{aligned} \int_T \omega_M \boldsymbol{\lambda} \cdot \nabla u^h \, d\mathbf{x} &= \beta_M^T \int_T \boldsymbol{\lambda} \cdot \nabla u^h \, d\mathbf{x} \\ &= \beta_M^T \left(\sum_{M \in T \cap \Gamma^-} k_M^+ g(M) + \sum_{M \in T, M \notin \Gamma^-} k_M^+ u_M \right) \\ &= \gamma_M^T \Phi_i^T. \end{aligned} \tag{31}$$

The first line is the Linearity Preserving formulation. This is detailed in the second line so that we see how to implement the boundary conditions. The last line is a rephrasing of the first one taking into account the fact that the LP scheme is constructed from a monotone first order scheme.

Hence, the inflow boundary conditions are simply implemented by setting $u = g$ at the vertices on Γ^-. Even though this is very simple, the scheme is still second order accurate. Note that (31) is generalized in a straightforward manner in the system case.

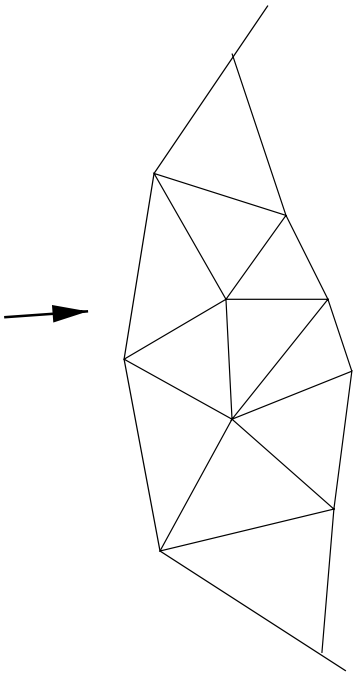

Fig. 4. Geometry near the inflow boundary Γ^-.

Still in the system case, in the Euler case more precisely, we still have to define the no-slip boundary conditions. This is done by imposing weakly the condition $\vec{u}\cdot\vec{n}=0$ on solid boundaries. As in [11] and several other references, we simply set $\vec{u}\cdot\vec{n}=0$ in the continuous flux, then linearly interpolate the pressure: this defines a numerical flux on the boundary. The residual interpretation of this flux is defined following the method of Section 2.3.2.1.

3. Numerical experiments

In this section, we present some numerical results obtained with the scheme (4) for two sets of problems: two scalar problems and two fluid mechanics ones. We particularly focus on the iterative convergence history.

3.1. Scalar problems

We first start with the two problems

$$\begin{aligned} &-y\frac{\partial u}{\partial x}+x\frac{\partial u}{\partial y}=0, \quad (x,y)\in[0,1]^2,\\ &u(x,0)=\begin{cases}-\sin\left(\pi\frac{x-0.7}{0.6}\right) & \text{if } x\in[0.1,0.7],\\ 0 & \text{else,}\end{cases}\end{aligned} \tag{32}$$

and

$$\begin{aligned} &\frac{1}{2}\frac{\partial u^2}{\partial x}+\frac{\partial u}{\partial y}=0,\\ &u(x,0)=1.5-x,\\ &u(0,y)=1.5,\\ &u(1,y)=-0.5\end{aligned} \tag{33}$$

with a CFL number of 0.5. Two schemes are evaluated: the scalar N-PSI scheme and the scheme constructed on the Lax–Friedrich scheme referred as the LxF-PSI scheme.

The L^2 convergence history is displayed in Fig. 5. Clearly, the convergence history of the LxF-PSI, after a good startup, becomes erratic. This is not the case of the N-PSI scheme, which has a very good behavior. These behaviors are characteristic of the schemes, whatever the CFL number.

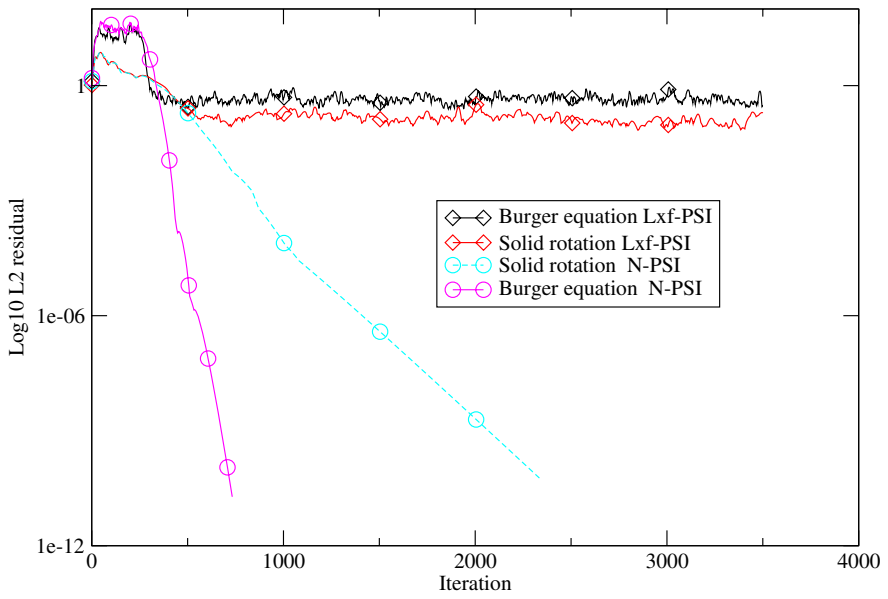

Fig. 5. Convergence history of the PSI and LxF-PSI schemes for (32) and (33) on the solid rotation and Burgers problems.

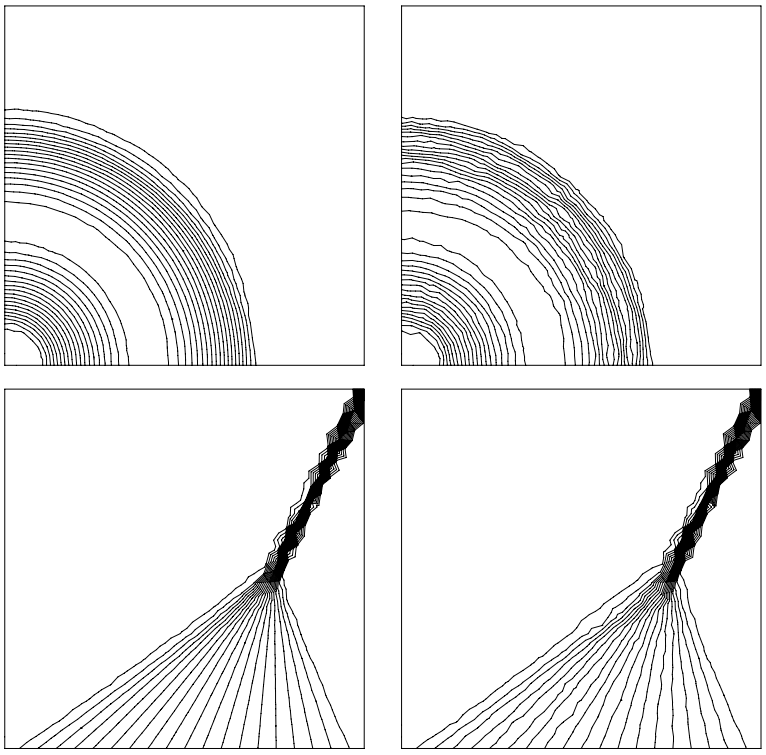

Fig. 6. Solutions for the N-PSI scheme – column (a) – and the Lxf-PSI scheme – column (b). Top: problem (32), bottom: problem (33).

If we look at the solution, see Fig. 6 we can observe wiggles in the solutions obtained by the Lax–Friedrich PSI scheme. These wiggles are not the manifestation of an instability: the scheme is perfectly stable in L^∞. In Fig. 7, we plot one cross-section for the rotation problem: there is no oscillation at all, but kinds of *plateau* develop. If we increase the resolution, this phenomena is amplified.

3.2. System problems

Consider the example of the Euler equations. The schemes are the system N-PSI scheme of [11] and the system LxF-PSI. These schemes are described in Section 2.3. We have done the same simulations for the Roe schemes (the standard one of [22] and the multidimensional one of Section 2.3.2.2), and we observe the same wiggly behavior and the same difficulties for the iterative convergence. Hence, all the results of this section are given for the N-PSI and LxF-PSI schemes.

To illustrate the erratic behaviors of the schemes, three test cases are considered which illustrate three flow regimes: subsonic, transonic and supersonic. The first example is a supersonic jet in a box $[0, 1]\times[0, 1]$. The inflow conditions are ($\gamma = 1.4$)

$$(\rho, u, v, p) = \begin{cases} (\gamma, 2.4, 0, 1) & \text{if } x > 0.5, \\ (\gamma, 4.4, 0, 1) & \text{else.} \end{cases} \tag{34}$$

The solution is everywhere supersonic and consists, from top to bottom, in a shock, followed by a contact line and then a fan. Since the flow is supersonic, there is no boundary condition problem: the iterative residuals (in the max norm and the L^2 norm) are not spoiled by any unclear effect of the boundary conditions implementation. They are displayed in Fig. 8.

These results show an erratic and very poor behavior of the residuals. If one looks at the Mach number isolines, displayed in Fig. 9, one can see some "wiggles" in each of the waves. Since the problem is self-similar

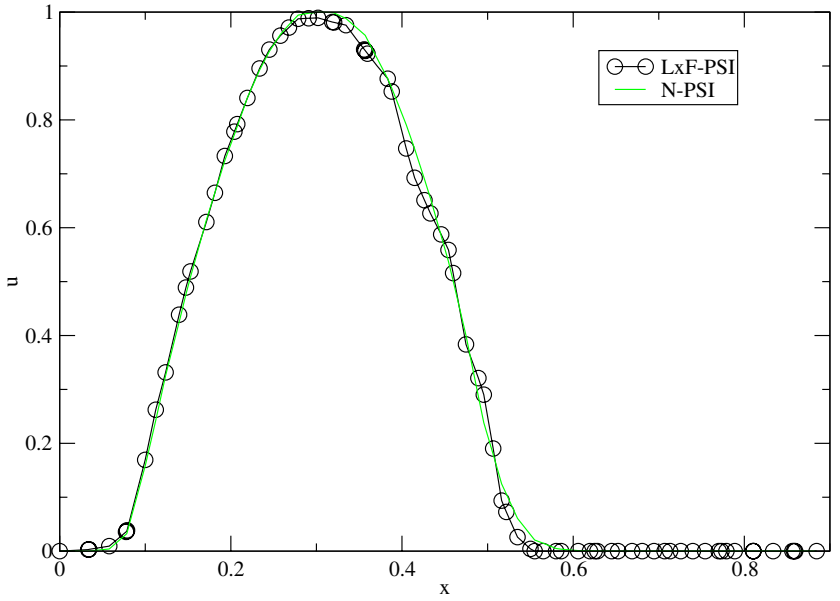

Fig. 7. Cross-section for the rotation problem (32). The PSI solution is plotted with plain lines, the LxF-PSI solution with circles.

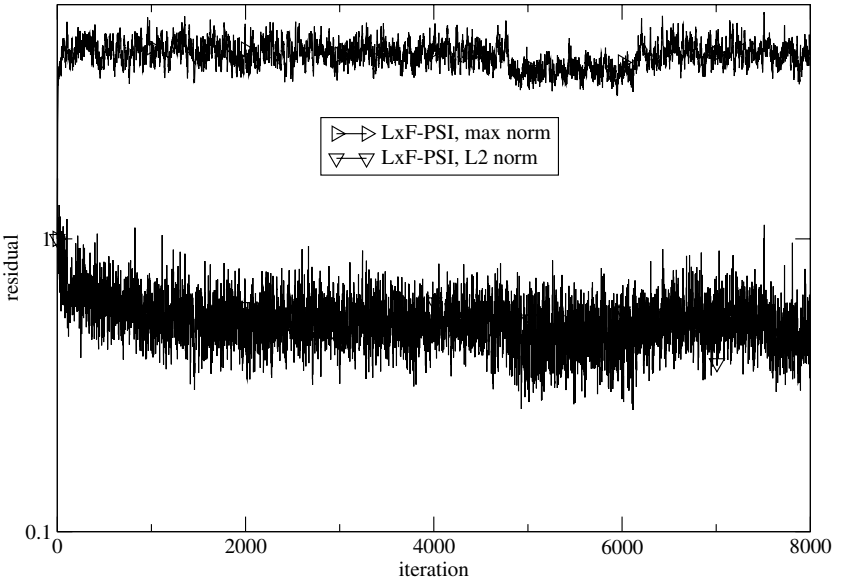

Fig. 8. Iterative residual (for density) in the max norm and the L^2 norm on the initial conditions (34).

with respect to (0, 0.5), the isolines should be straight lines focusing at (0, 0.5). The focusing is only approximate, and the isolines are far from straight lines. This behavior is independent of the CFL number. We have also noticed that the quality of results strongly depend on the variables that are in use for the second order construction (see Section 2.2.5). Here, the variables are the characteristic variables. If the conservative variables where used, the results would be even more wiggly.

A second case is considered. It is a fully subsonic flow over a sphere. The Mach number at infinity is $\mathscr{M}_\infty = 0.35$. This case is difficult and well documented, see for example [23]. The flow is steady and should be symmetric with respect to the vertical axis. Using the LxF-PSI scheme, we get the results displayed in Fig. 10.

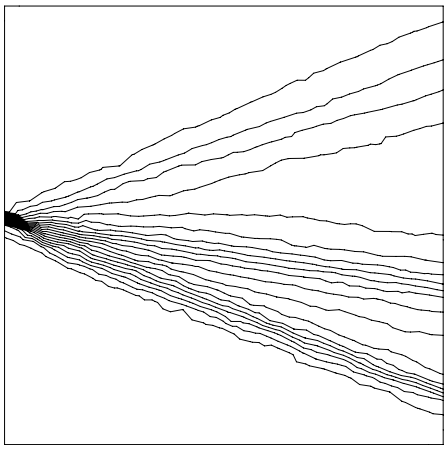

Fig. 9. Isolines of the Mach number for the problem (34) with the LxF-PSI scheme.

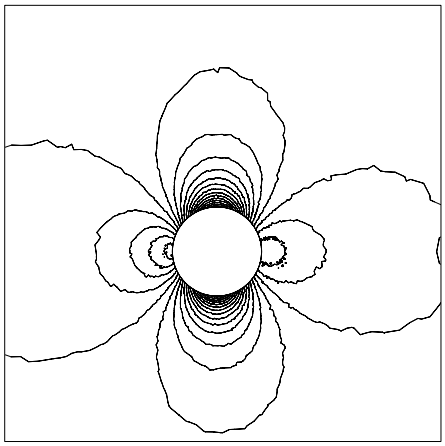

Fig. 10. Isolines of the pressure coefficient with the LxF-PSI scheme.

The flow is oscillatory. The convergence history is displayed in Fig. 11. As it can be seen, the convergence is very erratic too.

The last example is the NACA0012 case where the Mach number at infinity is $M = 0.85$, with 1° of incidence. The convergence history is similar to what happen in the previous case and is displayed in Fig. 12. Last, we display the Mach number isolines in Fig. 13. As before, the solution is wiggly on the smooth parts of the flow. However, the two shocks are very clean as it can be seen on the right of Fig. 13: there is no oscillation, and the shocks are resolved in one cell only.

3.3. Comments

The examples of this section shows that

- In the scalar case, the N-PSI behaves very well. In the case of the first order schemes, it is possible to exhibit analytically the dissipative mechanisms. In the case of the N-PSI scheme, this is much less clear. The main property of this scheme is its upwind nature. We conjecture that it is because of this upwind character that the N-PSI scheme has such a nice behavior. We provide argument in favor of this in the next section. We have also run the same cases with the scheme constructed from the first order (finite volume) Roe scheme where the PSI limiter of (9). The behavior of the scheme (quality of solutions,

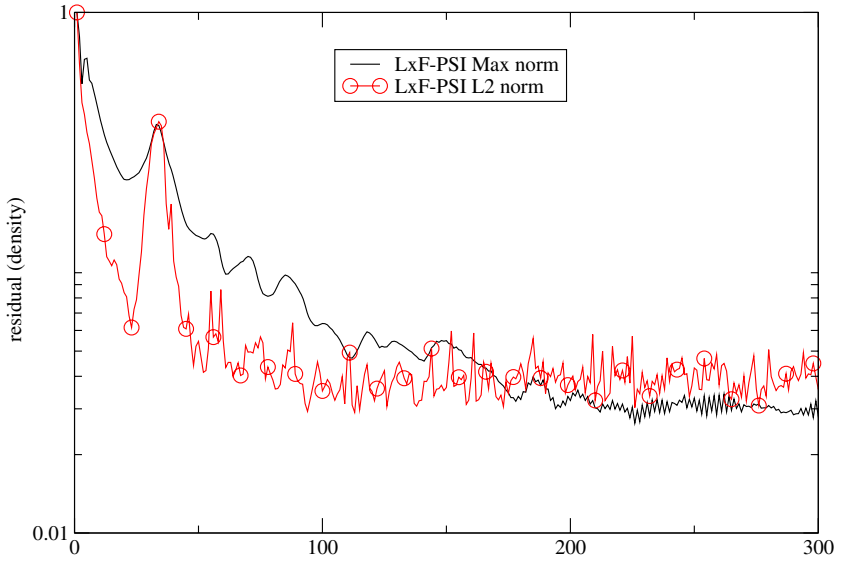

Fig. 11. Convergence history for the sphere problem, L^2 and max norm.

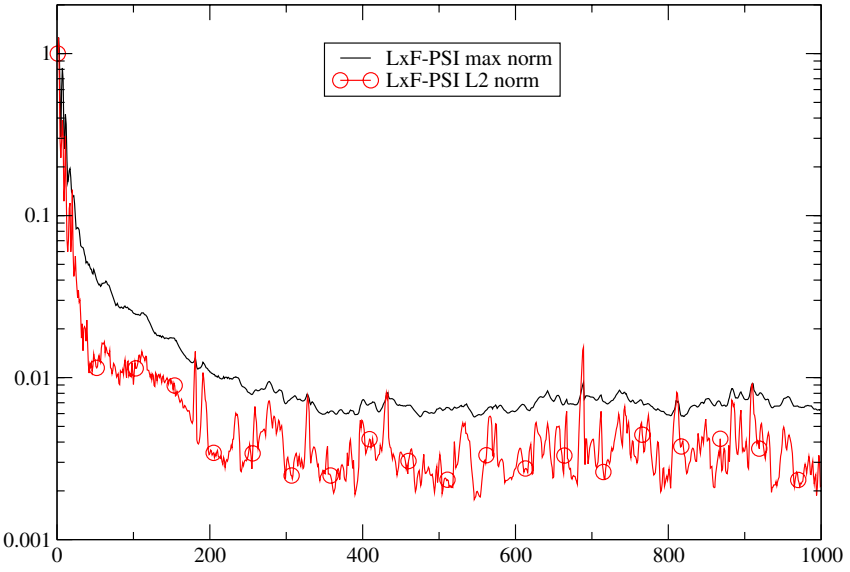

Fig. 12. Convergence history for the NACA0012 case, L^2 and max norm.

iterative convergence) is almost as good as for the N-PSI even though this scheme is not strictly speaking multidimensional upwind, see [24] for a discussion. It seems that starting from an upwind (or quasi upwind) scheme is a good point.

- However, in the system case, the PSI extension of the first order schemes, whatever they are, suffer from a degradation of the iterative convergence. The solution may look good (as for the N-PSI), but not the iterative convergence. Once more we observe that starting from an upwind scheme (the system N scheme here) is a good point, but this is not enough. We recall that the blended scheme (constructed from upwind schemes) presented in [5] or [1] have a very nice iterative convergence, but they are not robust enough.[2]

[2] In particular, their extension to unsteady problems is not satisfactory, this has motivated in [3] the introduction of the PSI extension of the system schemes.

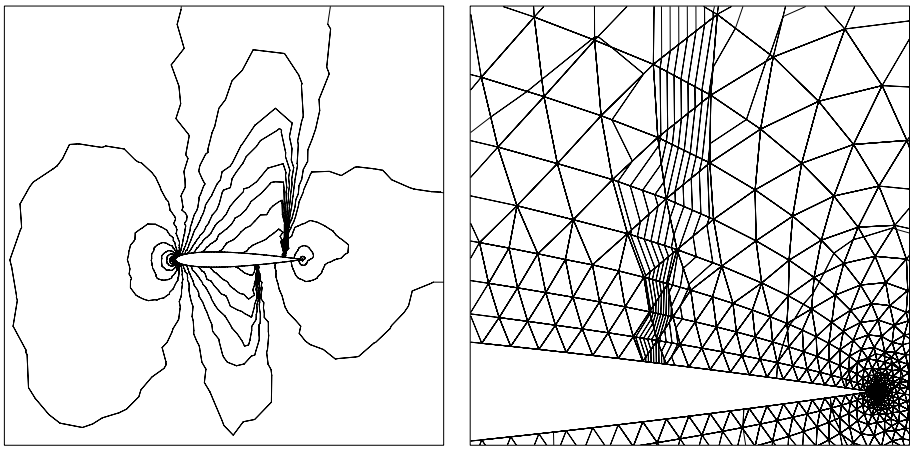

Fig. 13. Isolines of the Mach number. On the right, we zoom the solution near the upper shock. The mesh is also represented.

- All these example show that the problem is not a consequence of a wrong handling of discontinuities. In fact, the wiggles always occur in the *smooth* part of the flow. The discontinuities are always well handled.

In the next section, we provide some explanations of these strange behaviors and propose some modifications that do not destroy the non-oscillatory behavior of the schemes as well as their compactness.

4. How to remedy to convergence problems?

4.1. Analysis

We start again from the scalar version of (1) with $f(u) = \boldsymbol{\lambda} u$,

$$\begin{aligned} &\boldsymbol{\lambda}\cdot\nabla u = 0 \quad \text{in } \Omega,\\ &u = g \qquad\quad \text{on } \Gamma^-. \end{aligned} \tag{35}$$

From (6) and using Section 2.4, the schemes we consider write

$$\text{for any vertex not on } \Gamma^-, \quad a_{ii}u_i - \sum_{j\in\mathscr{V}(i)} a_{ij}u_j = f_i \tag{36}$$

with, see Fig. 14,

$$f_i = \begin{cases} 0 & \text{if } M_i \text{ is not connected to } \Gamma^-\\ \sum_{T\ni M_i}(\beta_i^T)\left[\sum_{\ell\in\Gamma^-\cap T}(k_\ell^T)^+ g(M_\ell)\right]. \end{cases} \tag{37}$$

The coefficients β_i^T are defined as in Section 2.2.5. The coefficients a_{ii} and a_{ij} will depend on the solution, and we have

$$\begin{aligned} &a_{ij} \geqslant 0,\\ &a_{ii} = \sum_{j\in\mathscr{V}(i)} a_{ij}, \end{aligned}$$

where $\mathscr{V}(i)$ denotes the set of nodes that are connected to M_i by an edge. For the ease of notations, (36) is written as

$$Au^h = f, \tag{38}$$

note that A may depend on u^h, and we denote by D the diagonal matrix $D = \mathrm{diag}(a_{ii})$.

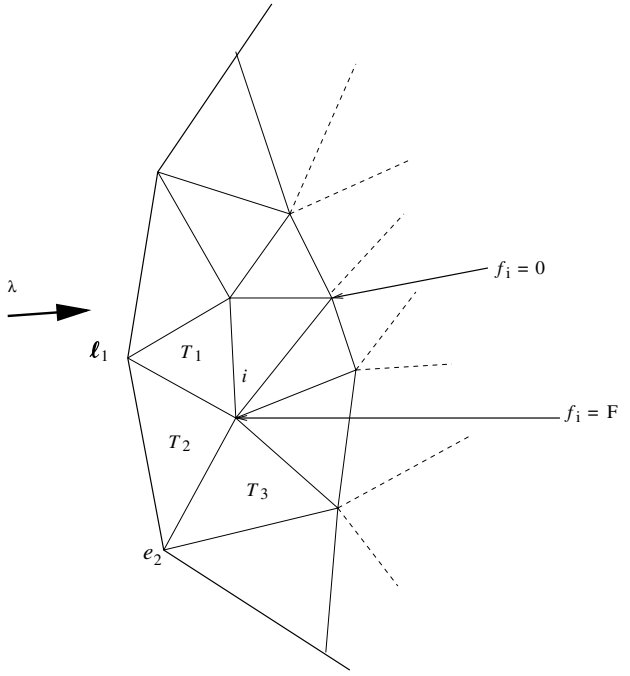

Fig. 14. Illustration of the relations (37). We have set $F = \sum_{T=T_1,T_2} \beta_{\ell_1}^T (k_{\ell_1}^T)^+ g(M_{\ell_1}) + \sum_{T=T_2,T_3} \beta_{\ell_2}^T (k_{\ell_2}^T)^+ g(M_{\ell_2})$.

The question is to see

1. whether the solution of (36) exists and is unique,
2. and whether there exists a norm and a constant independent of the mesh resolution such that $\|u\| \leqslant C\|u_0\|$,

as in the continuous case.

The answer to these questions is a difficult problem for which we can only provide qualitative answers.

4.1.1. Existence of a solution

The matrix $A(v)$ is homogeneous of degree 0 in v, as it can be seen from Section 2.2.5. Denoting by h the function $v \mapsto v - \omega(A(v)\cdot v - f)$, the scheme (4) writes $u = h(u)$. The sequence $u_{n+1} = h(u_n)$ converges if h admits a Lipschitz constant <1. Here, we have

$$h'(v) = \mathrm{Id} - \omega(A(v) + A'(v)\cdot v).$$

Since $v \mapsto A(v)$ is homogeneous of degree 0,[3] we have $A'(v)\cdot v = 0$. Hence a necessary and sufficient condition for the convergence of the scheme is that ω satisfies $\rho(\mathrm{Id} - \omega A(v)) < 1$. This condition is equivalent to the invertibility of $A(v)$, whatever v.

Let w such that $A(v)w \equiv Aw = 0$. For any i, we have

$$a_{ii}w_i = \sum_{j\neq i} a_{ij}w_j.$$

If i_0 is the index such that $\max|w_i| = |w_{i_0}|$. We get

$$a_{ii}|w_{i_0}| \leqslant \sum_j a_{ij}|w_j| \leqslant a_{ii} \max_{j\neq i_0} |w_j|.$$

[3] Here we assume unduly that the mappings $(x_1, x_2, x_3) \mapsto (\beta_1, \beta_2, \beta_3)$ introduced in Section 2.2.5 are smooth which is not completely true.

If for any i, $a_{ii} > 0$, we get that $|w_j| = |w_{i_0}|$ for all j. This corresponds to a check-board like mode: we can assume $w_i = \pm 1$. Then, by the Cauchy–Schwartz inequality, we have (since $a_{ij} \geqslant 0$)

$$a_{ii}^2 = a_{ii}^2 w_i^2 = \left(\sum_{j\neq i} a_{ij} w_j\right)^2 \leqslant \left(\sum_{j\neq i} a_{ij}\right)\left(\sum_{j\neq i} a_{ij} w_j^2\right) \leqslant a_{ii}\left(\sum_{j\neq i} a_{ij}\right) \leqslant a_{ii}^2.$$

In other words,

$$\left(\sum_{j\neq i} a_{ij} w_j\right)^2 = \left(\sum_{j\neq i} a_{ij}\right)\left(\sum_{j\neq i} a_{ij} w_j^2\right)$$

and then, by the Cauchy–Schwartz inequality again, there exists μ such that

$$\text{for any } j \neq i, \quad \sqrt{\frac{a_{ij}}{a_{ij}}} = \mu w_j > 0.$$

We can assume that $w_j > 0$, so $w_j = 1$ and then $w_i = 1$: the only spurious mode is $(1, 1, \ldots, 1)$. This provides information on the structure of the matrix when it is not invertible: A is not invertible if and only if one of the two conditions hold:

1. there exists on index for which $a_{ii} = 0$ in which case $a_{ij} = 0$ whatever j,
2. whatever i, $a_{ii} = \sum_{j\neq i} a_{ij}$.

For any vertex i that has no common edge with the inflow boundary, we know that $a_{ii} = \sum_{j\neq i} a_{ij}$ because the stencil of the scheme at i is the set of its immediate neighbors. Hence the necessary and sufficient conditions for the invertibility of A are

1. for any i, $a_{ii} > 0$,
2. whatever i that has a common edge with the inflow boundary, $a_{ii} > \sum_{j\neq i} a_{ij}$.

In the case of an upwind scheme, such as the N scheme and the N-PSI scheme, we can check that these conditions are verified. The structure of the matrix $A := (a_{ij})_{i,j}$, after a suitable numbering of the vertices (Fig. 15),[4] is

$$A = \begin{pmatrix} A_{11} & 0 & 0 & \cdots & 0 \\ A_{21} & A_{22} & & \cdots & 0 \\ 0 & A_{23} & A_{33} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & A_{n-1\,n} & A_{nn} \end{pmatrix}. \tag{39}$$

The matrices A_{lp} are block matrices, and the indices ξ and ν correspond to line indices. Greek indices are used to denote line indices and Latin symbols for vertices: for $A = (a_{ij})_{i,j}$, the indices i and j correspond to vertex indices.

For a regular mesh of n_s vertices, there is $\mathcal{O}(\sqrt{n_s})$ lines: the matrix A is $n_s \times n_s$, while there is only $O(\sqrt{n_s})$ non-zero blocks in (39). If D is the diagonal matrix where $d_{ii} = a_{ii}$, we notice that $D(v)$ is invertible whatever v and $d_{ii} = a_{ii} = \mathcal{O}(h)$.

In fact, for any vertex M_i, there is one triangle for which the N scheme is one target: the case of internal vertices is clear, the case of boundary vertices also hold true because of the very definition of the inflow boundary condition. Denote by T^i_{up} this triangle. From Sections 2.3.1.1 and 2.2.5, and relations (6) and (7), since $c_{ij}^T \geqslant 0$, we first have

[4] The vertices are ordered by lines of increasing arrival times from the inflow boundary, see Fig. 15.

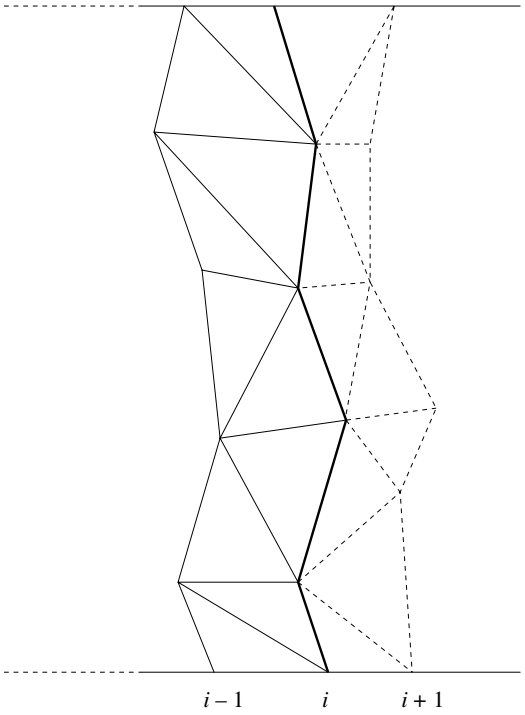

Fig. 15. Example of numbering of the mesh by arrival time.

$$a_{ii} \geqslant \sum_{j \in T^i_{\text{up}}} c_{ij}^{T^i_{\text{up}}}. \tag{40}$$

This is independent of v (which plays a role in the evaluation of β_i, but in the case of a one target triangle, $\beta_i = 1$.) Then:

- For the N scheme, since $\Phi_i^{T^i_{\text{up}}} = \Phi^{T^i_{\text{up}}}$ (this is the one target property), we have $c_{ij}^{T^i_{\text{up}}} = k_i$. Then $k_i > 0$: this is the one target property once more, and then, $a_{ii} = \mathcal{O}(h)$.
- For the N-PSI scheme, we have $\Phi_i^{T^i_{\text{up}}} = \Phi^{T^i_{\text{up}}}$ thanks to the one target property, once more.

This proves that D is invertible.

Next, we consider $D^{-1}A$. This matrix has negative off diagonal coefficients, and clearly we can write the block diagonal matrix $D^{-1}A_{ii}$ as $D^{-1}A_{ii} = \mathrm{Id} - B_{ii}$, where B_{ii} is strictly diagonal dominant. Thus $D^{-1}A_{ii}$ is invertible. This shows that $A(v)$ is invertible whatever v.

4.1.2. Continuous dependence with respect to the data

The sequence $u_{n+1} = h(u_n)$ writes for any mesh point M_i,

$$(u_{n+1})_i = (u_n)_i - \omega \left\{ \sum_{j \in \mathcal{V}(i)} a_{ij} \Big((u_n)_i - (u_n)_j \Big) - f_i \right\},$$

where $a_{ij} \geqslant 0$. Thus, from the definition of f_i in Section 2.2.3 (in particular the scheme is constructed from a monotone first order scheme) and under the condition $a_{ii}\omega < 1$, we have

$$\max_i |(u_{n+1})_i| \leqslant \max_{M_i \in \partial\Gamma^-} |g_i|.$$

If the sequence converges, we have the stability inequality

$$\max_i |u_i| \leqslant \max_{M_i \in \partial\Gamma^-} |g_i|.$$

This result is well known.

The second remark is that under the same assumptions, we can get an error estimate. If $\pi^h u$ is an interpolant of the true solution of (35) and if $\pi^h g$ is a piecewise linear interpolant of g on Γ^-, since the scheme is LP, we have, setting $e^h = u^h - \pi^h u$,

$$a_{ii} e_i - \sum_j a_{ij} e_j = f'_i,$$

where f' is defined as f in (37) with g replaced by $g - \pi^h g$. This result can be seen from (30). Since $\max_{M_i \in \Gamma^-} |f'_i| = \mathcal{O}(h^2)$ (this is an interpolant), we get

$$\max_{i \notin \Gamma^-} |e_i| \leqslant Ch^2 \tag{41}$$

with the constant C independent of g.

4.1.3. Comments

We consider a family of regular triangulations and the problem

$$\begin{aligned} &\boldsymbol{\lambda} \cdot \nabla u = f, \quad && x \in \Omega, \\ &u = g, && x \in \partial\Gamma^-. \end{aligned}$$

Since $\boldsymbol{\lambda}$ is non-zero, we can order the vertices as we have done for the N scheme. This defines, for any vertex, the downwind nodes.

Our conjecture for the convergence of the iterative method is that this method converges if the following two conditions are true:

1. there exists $\alpha > 0$ independent of the considered triangulation in the family such that whatever i, $a_{ii} \geqslant \alpha h$;
2. for any i, $a_{ii} > \sum_{j \in \mathcal{V}(i),\ \text{non-downwind nodes}\ a_{ij}}$ as in (40).

The second condition enables a coupling between the vertex M_i and its downwind nodes, so that information can propagate.

4.1.4. Geometrical interpretation

It may be interesting to visualize the way the PSI versions of a first order scheme behaves. Consider Fig. 16. The arrows represent the non-vanishing distribution coefficients, i.e. they indicate the vertices of T where "something" is sent: $x_i = \Phi_i / \Phi^T \neq 0$. In the case of the N scheme, these vertices are always downwind, this is no longer true in the case of the LxF scheme where a priori non-zero residual are sent at each vertex.

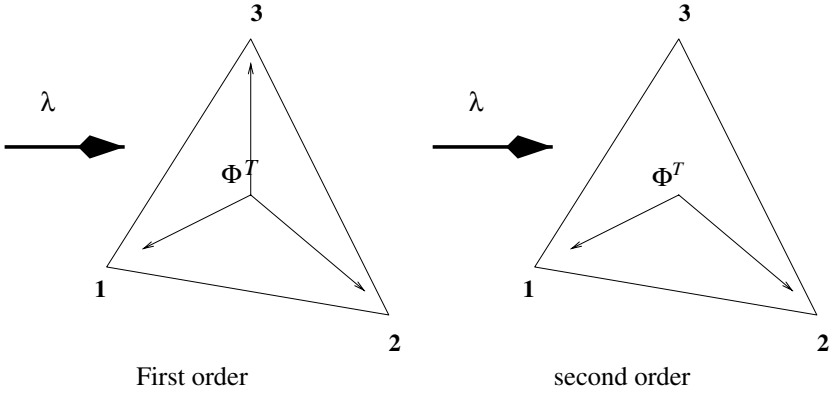

Fig. 16. Geometrical illustration of the non-vanishing distribution coefficients.

Then we apply the mapping $(x_1, x_2, x_3) \mapsto (\beta_1, \beta_2, \beta_3)$ as in Section 2.2.3. In the case of the PSI limiter, if one of the x_i's is outside of $[0, 1]$, necessarily one of the β_j is set to zero. This is done according the signs of the distribution coefficients x_i, and *not* using any consideration about the upwind or downwind nature of the triangle vertices. In other word, as in Fig. 16, "something" can be sent to a downwind node. This has a destabilizing effect which is corrected by the fact that the β_i's are defined in order to guaranty the local L^∞ bounds. We have no control on the coefficient a_{ii} and it may be that all downwind coefficients are set to zero. This is precisely this destabilizing character that has to be corrected.

4.2. Two solutions

From the previous analysis, we conjecture that the wiggles are consequence of a bad structure of the A matrix. We conjecture that the diagonal coefficients must satisfy $a_{ii} \geqslant \delta h$ for $\delta > 0$ uniform, and that the matrix A must be uniformly invertible in some norm. In the previous section, we have stressed on the maximum norm. In this section, we propose two methods to enforce these two properties.

4.2.1. First solution

This solution has been imagined by M. Mezine in his thesis [25]. It works only for scalar problems. Once more we consider the problem

$$\begin{aligned} &\boldsymbol{\lambda} \cdot \nabla u = f, \quad x \in \Omega, \\ &u = g, \qquad\quad x \in \partial\Gamma^-. \end{aligned}$$

He starts from a monotone scheme to which we apply the limitation technique of Section 2.2.5. Any any vertex M_i (including the boundaries provided they are non-characteristic) has a single upwind triangle for which $k_i > 0$ and $k_j < 0$ for the two other vertices. For this triangle T^{up}, we modify the limited residual by setting

$$\Phi_i^{T^{\text{up}}} = \Phi^{T^{\text{up}}}, \quad \Phi_j^{T^{\text{up}}} = 0 \qquad \text{for the other vertices.}$$

Clearly, $a_{ii} \geqslant k_i \geqslant \alpha h$ and the second condition of our conjecture also holds.

In the case of a nonlinear problem, the same properties about a_{ii} and the inequality (40) holds except maybe when $\nabla_u f \simeq 0$ around M_i.

4.2.2. Second solution

The main problem of the previous correction is that it can apply *only* to scalar problems because it deeply relies on the study of the signs of the inflow parameters k_j. Because of that, we present now a second solution that applies to systems. The price to pay is to lose the maximum norm property. Here, we work with the energy norm which is more tractable for systems.

We start again by a scheme of the type

$$u^{n+1} = u^n - \omega(Au^n - f).$$

The scheme satisfies $r = \|\mathrm{Id} - \omega A\|_{L^2} < 1$ with $\omega > 0$ if for any $v \in \mathbb{R}^n$, we have

$$\|(\mathrm{Id} - \omega A)v\|^2 = \|v\|^2 - 2\omega\langle Av, v\rangle + \omega^2\|Av\|^2 \leqslant r\|v\|^2.$$

Since $\omega > 0$, there must exist a positive root to

$$-2\langle Av, v\rangle + \omega\|Av\|^2 \leqslant 0.$$

This is possible only if

$$\langle Av, v\rangle > 0;$$

this is the well-known dissipation condition.

In the present case, the iterative scheme is

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|C_i|}\sum_{T \ni i} \Phi_i^T$$

and we assume $\Phi_i^T = \beta_i^T \Phi^T$ with β_i^T bounded.

The natural scalar product is

$$\langle u, v\rangle := \sum_i |C_i| u_i v_i$$

and then

$$\langle u^{n+1}, u^{n+1}\rangle = \langle u^n, u^n\rangle - \Delta t \sum_i u_i^n \left(\sum_{T\ni i} \Phi_i^T\right) + \Delta t^2 \sum_i \frac{\left(\sum_{T\ni i} \Phi_i^T\right)^2}{|C_i|}.$$

The dissipation condition writes

$$\sum_i u_i^n \left(\sum_{T\ni i} \Phi_i^T\right) > 0,$$

that is (forgetting the temporal superscript)

$$\sum_T \left(\sum_{j\in T} \beta_j^T u_j\right) \int_T \boldsymbol{\lambda}\cdot\nabla u \,\mathrm{d}\mathbf{x} > 0$$

or equivalently

$$\int_\Omega \ell(u)\boldsymbol{\lambda}\cdot\nabla u \,\mathrm{d}\mathbf{x} \geqslant 0$$

with

$$\ell(u)_{|T} = \sum_{j\in T} \beta_j^T u_j.$$

Assume now that the scheme writes as

$$\Phi_i^T = \beta_i^T \Phi^T + \alpha \int_T (\boldsymbol{\lambda}\cdot\nabla\varphi_i)(\boldsymbol{\lambda}\cdot\nabla u)\,\mathrm{d}\mathbf{x},$$

where α is a real parameter and φ_i is the linear basis function associated to the vertex M_i. The scheme is dissipative if

$$\int_\Omega \ell(u)\boldsymbol{\lambda}\cdot\nabla u \,\mathrm{d}\mathbf{x} + \alpha \int_\Omega (\boldsymbol{\lambda}\cdot\nabla u)^2 \,\mathrm{d}\mathbf{x} \geqslant 0$$

and we look for α to have this property.

First, we rewrite the original scheme as

$$\Phi_i^T = \frac{\Phi}{3} + h\int_T (\vec{\xi}_T\cdot\nabla\varphi_i)(\boldsymbol{\lambda}\cdot\nabla u)\,\mathrm{d}\mathbf{x}$$

with

$$\vec{\xi}_T = \sum_j \left(\beta_j - \frac{1}{3}\right)\overrightarrow{GM_j},$$

where G is the centroid of T. Here, we look for $\alpha = \theta h$ such that

$$\int_\Omega (\vec{\xi}\cdot\nabla u)(\boldsymbol{\lambda}\cdot\nabla u)\,\mathrm{d}\mathbf{x} + \theta\int_\Omega (\boldsymbol{\lambda}\cdot\nabla u)^2\,\mathrm{d}\mathbf{x} \geqslant 0. \tag{42}$$

It is known that $u\mapsto\sqrt{\int_\Omega(\boldsymbol{\lambda}\cdot\nabla u)^2\,\mathrm{d}\mathbf{x}}$ defines a norm on the functions that vanish on the inflow boundary. Since the space of linear functions that vanish on Γ^- is finite dimensional, there exists θ_h^0 such that (42) is true. We take $\theta_h > \theta_h^0$.

Thus we modify the original residual (8) into

$$\Phi_i = \beta_i \Phi + h\theta_h \int_T (\boldsymbol{\lambda}\cdot\nabla\varphi_i)(\boldsymbol{\lambda}\cdot\nabla u)\,\mathrm{d}\mathbf{x} = \beta_i\Phi + \frac{\theta_h}{h}k_i\Phi. \tag{43}$$

In that case, it is clear that the coefficients a_{ii} as in (6) and (7) satisfies

$$a_{ii} \geqslant \delta h$$

for $\delta > 0$ independent of h. The uniform invertibility comes from the fact that

$$\int_{\Gamma^-} \frac{u^2}{2}\boldsymbol{\lambda}\cdot\vec{n}\,\mathrm{d}\hat{\sigma}\Omega + (\theta_h - \theta_h^0)h\int_\Omega (\boldsymbol{\lambda}\cdot\nabla u)^2\,\mathrm{d}\mathbf{x}$$

defines a norm, as for the standard streamline diffusion method.

Several choices of θ will be considered in the next section. A priori, we let θ_h depends on the solution itself, $\theta_h \equiv \theta(u^h)$. Unfortunately, the monotonicity preserving property is formally lost. In the numerical applications, we see that the convergence properties of the scheme are good. The monotonicity properties of the original scheme are quasi preserved. This can be improved by better choices of θ as we see later. In particular, we look for $\theta(\mathbf{u}^h)$ such that $\theta(\mathbf{u}^h) \equiv 0$ in discontinuities.

In the system case, the relation (14) is modified into

$$\Phi_i = \mathbf{B}_i\Phi + \boldsymbol{\Theta}(\mathbf{u}^h)h^{-1}\mathbf{K}_i\boldsymbol{\Theta}(\mathbf{u}^h)\Phi. \tag{44}$$

to respect symmetry. Here $\boldsymbol{\Theta}$ is chosen to be a diagonal matrix, several choices will be discussed in the next section which are all proportional to the identity matrix. Better choices could certainly be investigated.

5. Numerical experiments revisited

We rerun the cases of Section 3 with the schemes (43) and (44). In addition and for scalar problems, we display the results of the scheme constructed from the Lax–Friedrich-PSI scheme with the modification due to M. Mezine.

5.1. Scalar case

Different choices of θ are considered namely $\theta = \theta_j$ defined by

- $\theta_1 = 1$.
- For the Burgers equation we let

$$\theta_2 = \begin{cases} 1 & \text{if the } y\text{-component of centroid of } T \text{ is } > 0.5, \\ 0 & \text{else.} \end{cases}$$

  By doing so, we want to check whether the convergence problem is really located in the smooth part of the flow, since we know that the discontinuity is located at $y \geqslant 0.5$.
- Here the idea is identify the discontinuities. We know that for a smooth function, $\boldsymbol{\lambda}\cdot\nabla u^h = \mathcal{O}(1)$ and when ∇u is not discontinuous, $\nabla u/u = \mathcal{O}(h^{-1})$, so we choose

$$\theta_4 = \min\left(1, \frac{1}{\dfrac{|\Phi^T|}{\bar{u}h^2}}\right).$$

In that case we see that

$$\text{if } \boldsymbol{\lambda}\cdot\nabla u^h/\bar{u}^h = \mathcal{O}(1), \quad \theta_4 = \min\left(1, \frac{h}{\mathcal{O}(h)+h}\right) \equiv 1,$$

$$\text{if } \boldsymbol{\lambda}\cdot\nabla u^h/\bar{u} = \mathcal{O}(h^{-1}), \quad \theta_4 = \min\left(1, \frac{h}{\mathcal{O}(h^{-2})+h}\right) = \mathcal{O}(h).$$

In practical implementations, we have chosen

$$\theta_4 = \min\left(1, \frac{1}{\frac{|\Phi^T|}{\bar{u}h^2} + \varepsilon}\right)$$

with $\varepsilon = 10^{-10}$.

In the following, we add the suffix -D to the name of the scheme to indicate that the term (43) with a choice of θ that is also indicated. For example, the N-PSI scheme extends to N-PSI-D scheme.

Comparing Figs. 6 and 17, we see that the wiggles problem is cured whatever the choice of θ. We also see that Mezine's trick also permits to solve it: this is an indication of our conjecture about the origin of the problem (diagonal coefficients too small) has some content. We also see that there is no undershoot and overshoot problem, and last that the choice $\theta = \theta_4$ leads to slightly more dissipative results than $\theta = \theta_1$: the isolines are good approximations of circles and if one looks at the farthest from the origin, the approximation is better in the case $\theta = \theta_1$ (and in the case of Mezine's trick) than for $\theta = \theta_4$. The new schemes are not strictly positivity preserving: the minimum should be 0, it is in fact 0 for the LxF-PSI, N-PSI and LxF-PSI with Mezine's trick, it is -0.001312 for the choice $\theta = \theta_1$ and -0.00057 if $\theta = \theta_4$. This also confirm the fact that $\theta = \theta_4$ leads to a more dissipative (or more "positive") scheme. Fig. 18 gives the convergence histories for the various schemes. In each case, the convergence is smooth, this has to be compared with Fig. 5. In Fig. 19 we have displayed the L^2 error obtained by the LxF-PSI and LxF-PSI-D schemes for successive meshes and for the rotation problem. The meshes are obtained from an initial coarse one and successfully refined by adding the mid-edge points. The LxF-PSI is only first order while the LxF-PSI-D is clearly second order accurate.

In the case of the Burgers problem, compare now Figs. 6 and 20. The same conclusions hold: no more wiggle, a clean behavior in the discontinuity. The LxF-PSI with Mezine's trick is monotone, the LxF-PSI with

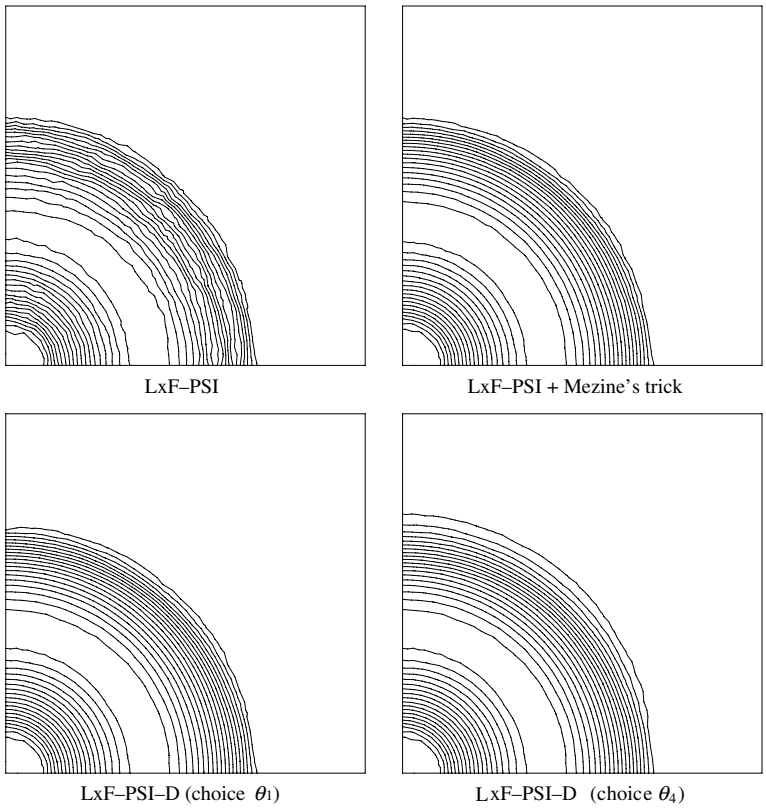

Fig. 17. Rotation problem. The baseline-first order scheme is the LxF scheme. The solution without dissipation, with Mezine's trick and the choices θ_1 and θ_4 are displayed. These results have to be compared to the N-PSI scheme displayed in Fig. 6.

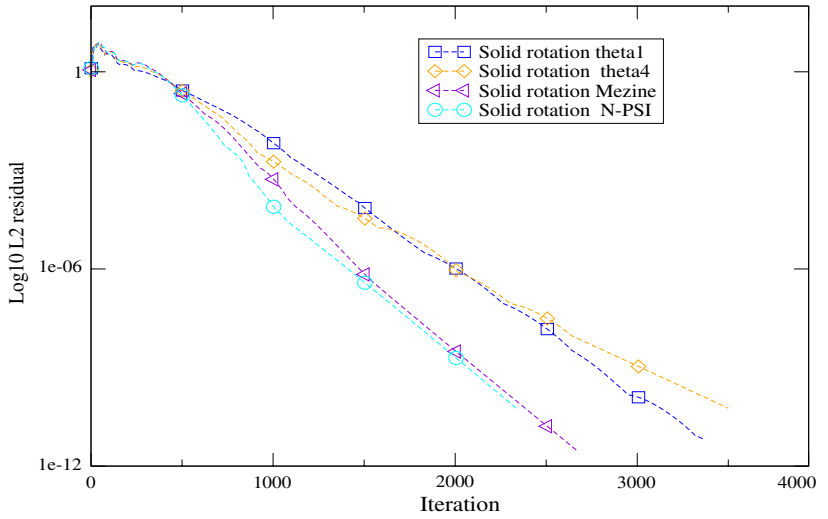

Fig. 18. Convergence history (L^2 norm) for the LxF-PSI with dissipation in the solid rotation problem: choices $\theta = \theta_1$, $\theta = \theta_4$.

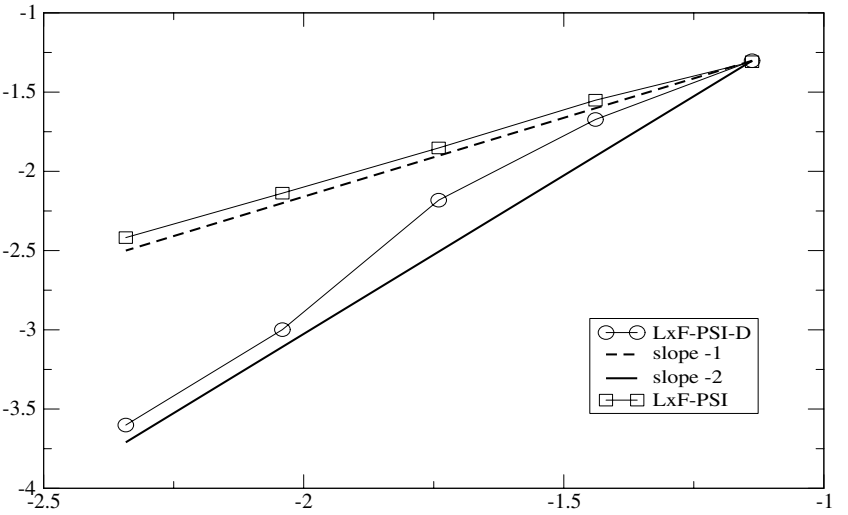

Fig. 19. Error plot (exact solution vs computed solution) in the L^2 norm. This is done for the LxF-PSI and LxF-PSI-D scheme. The slopes -1 and -2 are also represented.

$\theta = \theta_1$ or $\theta = \theta_4$ are not exactly monotone, since the solution belongs to $[-0.5, 1.5]$ when $\theta = \theta_1$ and $[-0.5052, 1.5]$ in the second case. The shock is slightly enlarged in the case $\theta = \theta_1$ compared to the other cases. In Fig. 21, we see that the choice $\theta = \theta_2$, though non-smooth, does not prevent the iterative convergence to be excellent. This fact might be surprising at first glance, but in a shock one can see from (9) that in fact $\Phi_i^*/\Phi_i = \mathcal{O}(1)$: the properties of the matrix A in (38) are not modified by the limiter.

The best compromise between accuracy and stability seems to be the choice $\theta = \theta_4$. It is surprising to see that there is *no* major difference between the N-PSI scheme (which provides the best results) and the LxF-PSI-D schemes whatever the choice of θ, and one has to remember that the LxF scheme is *very* dissipative! In the rest of the text, we choose the parameter $\theta = \theta_4$ and its generalization for the system cases.

5.2. System case

We test our technique on several test cases. We start from several first order schemes

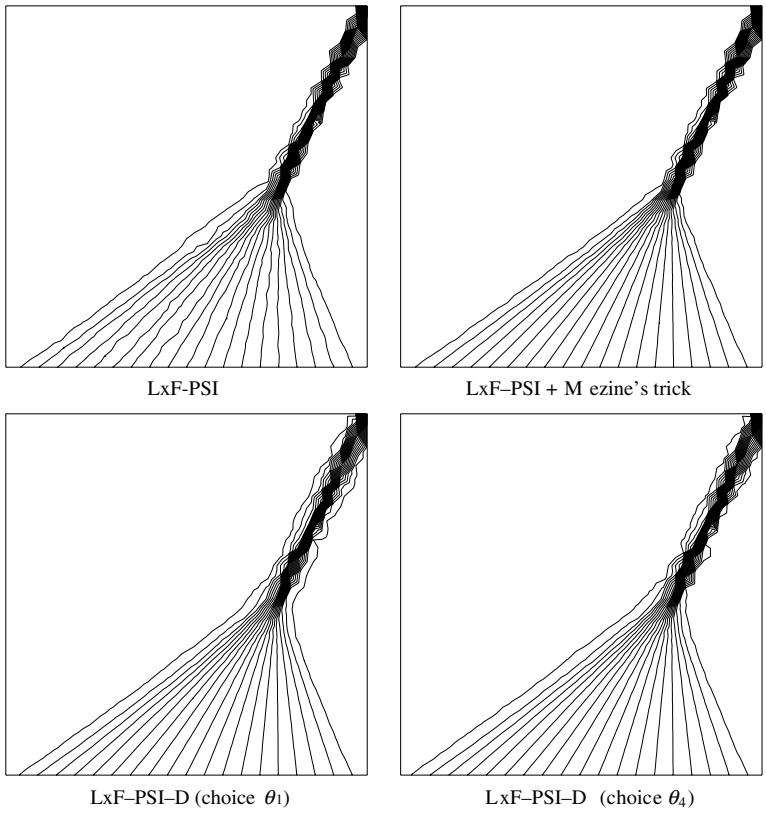

Fig. 20. Burgers problem. The baseline-first order scheme is the LxF scheme. The solution without dissipation, with Mezine's trick and the choices θ_1 and θ_4 are displayed. These results have to be compared to the N-PSI scheme displayed in Fig. 6.

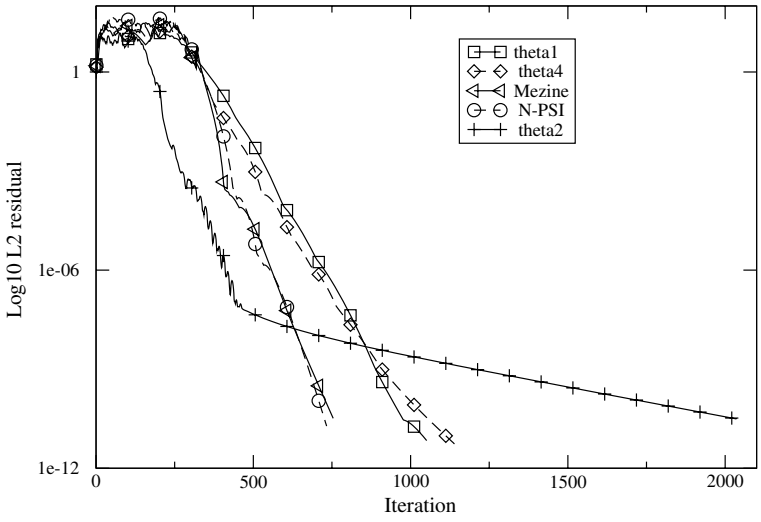

Fig. 21. Convergence history (L^2 norm) for the LxF-PSI with dissipation for the Burgers equation: choices $\theta = \theta_1$, $\theta = \theta_4$, $\theta = \theta_2$ and Mezine's trick.

- A Lax–Friedrich type scheme,

$$\Phi_i^T = \frac{1}{3}\left(\int_T \operatorname{div}\mathbf{f}(\mathbf{u}^h)\,\mathrm{d}\mathbf{x} + \alpha_T \sum_{j\neq i}(\mathbf{u}_i - \mathbf{u}_j)\right),$$

where $\mathbf{u}^h$ is evaluated via the Roe'Z–parameter vector

$$Z = \sqrt{\rho}(1, \mathbf{u}, H)^T$$

which is linearly interpolated in T and $\alpha_T = \max(\rho(K_1), \rho(K_2), \rho(K_3))$ where $\rho(A)$ represents the spectral radius of the matrix A.

- The system N scheme of van der Weide and Deconinck [20],

$$\Phi_i = \mathbf{K}_i^+(\mathbf{u}_i - \tilde{\mathbf{u}})$$

and

$$\tilde{\mathbf{u}} = \left(\sum_{j=1}^{3} \mathbf{K}_j^+\right)^{-1} \left(\sum_{j=1}^{3} \mathbf{K}_j^+ \mathbf{u}_j - \int_T \operatorname{div} \mathbf{f}(\mathbf{u}^h)\, d\mathbf{x}\right).$$

In [26], we show that $\sum_{j=1}^{3} \mathbf{K}_j^+$ is invertible except at stagnation points. However, the matrices

$$\left(\sum_{j=1}^{3} \mathbf{K}_j^+\right)^{-1} \mathbf{K}_l^{\pm}$$

always have a meaning, see this reference for more details.

- Roe's finite volume scheme denoted by Roe.
- Roe's multi D denoted by Roe2. It is defined in Section 2.3.2.2.

For each first order scheme, we construct the Linearity Preserving scheme as in [26]: we consider $\vec{n} = \vec{\mathbf{u}}/\|\vec{\mathbf{u}}\|$ and the average Jacobian matrices evaluated at the average state $\bar{\mathbf{u}}$ defined by the Roe average. This choice is not essential since other average states can be used. Then, we introduce the right eigenvectors $(\mathbf{r}_p)_{p=1,4}$ of $\nabla_{\mathbf{u}}\mathbf{f}(\bar{\mathbf{u}})$ and the corresponding left eigenvectors $(\ell_p)_{p=1,4}$. In this choice, the first eigenvector is associated to the entropy field: if $(\bar{u}, \bar{v})$ is the velocity field defined by $\bar{\mathbf{u}}$, we set

$$\mathbf{r}_1 = \begin{pmatrix} 1 \\ \bar{u} \\ \bar{v} \\ \frac{\bar{u}^2+\bar{v}^2}{2} \end{pmatrix}.$$

If $\bar{c}$ is the average sound speed defined by $\bar{\mathbf{u}}$, the corresponding left eigenvector is defined by its action on a state (A, B, C, D) by

$$\ell_0[(A, B, C, D)] = A - \frac{\gamma - 1}{\bar{c}^2}\left(D - \bar{u}B - \bar{v}C + \frac{\bar{u}^2 + \bar{v}^2}{2}A\right). \tag{45}$$

The first term of the right hand side of (45) corresponds to the density, the term $D - \bar{u}B - \bar{v}C + \frac{\bar{u}^2+\bar{v}^2}{2}A$ corresponds to the pressure variation.

This leads to the LxF-PSI, N-PSI, Roe-PSI and Roe2-PSI schemes. Last, we add the additional dissipation

$$\theta h \int_T \mathbf{K}_i \left(\overline{\mathbf{A}}\frac{\partial \mathbf{u}^h}{\partial x} + \overline{\mathbf{B}}\frac{\partial \mathbf{u}^h}{\partial y}\right) d\mathbf{x}$$

where we chose

$$\theta = \min\left(1, \frac{1}{\frac{|\varphi^T|}{|T|} + \varepsilon}\right).$$

Here $\varepsilon = 10^{-10}$ and

$$\varphi^T = \ell_0(\Phi_T)$$

The idea is that φ^T is an approximation of $\mathscr{S} = \rho(\bar{u}\frac{\partial s}{\partial x} + \bar{v}\frac{\partial s}{\partial y})$: when the flow is smooth, $\mathscr{S} \simeq 0$ while when a discontinuity exists, $\mathscr{S}/(\rho\sqrt{\bar{u}^2+\bar{v}^2}s) \simeq 1$. Thus, it the first case, $\theta \simeq 1$ and in the second one, $\theta \simeq 0$. We are interested in *steady* problems in this paper. In order to improve the efficiency, the schemes are implicit. Formally, instead of solving

$$F_i(\mathbf{u}) = 0, \quad i = 1, \ldots n_s,$$

we would solve

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \omega_i F_i(\mathbf{u}^{n+1}), \quad i = 1, \ldots n_s,$$

which is too complex. Instead, we adopt a standard linearized procedure

$$\left[\left(\mathrm{Id} + \omega F'_{\mathbf{u}}(\mathbf{u}^n)\right)(\mathbf{u}^{n+1} - \mathbf{u}^n)\right]_i = \mathbf{u}_i^n - \omega F_i(\mathbf{u}^n).$$

The evaluation of the Jacobian $F'_{\mathbf{u}}(\mathbf{u}^n)$ is too complex. Following a standard procedure, $F'(\mathbf{u}^n)$ is approximated by the Jacobian of the first order scheme.

In each case, we present the solution of the original PSI schemes and the PSI-D ones. We also display the convergence histories.

5.2.1. NACA012 airfoil

This problem is the same as in Section 3.2: the inflow Mach number is $\mathscr{M}_\infty = 0.85$ and the angle of incidence is 1°. We compare the N-PSI, LxF-PSI, N-PSI-D and LxF-PSI-D schemes. We have tested the different Roe schemes on this case, the results are similar and are not displayed here.

The Fig. 22 displays the density isolines. The shock waves are clearly non-oscillatory in all cases while, as expected, the LxF-PSI schemes density isolines behave badly in the smooth part of the flow. With the new

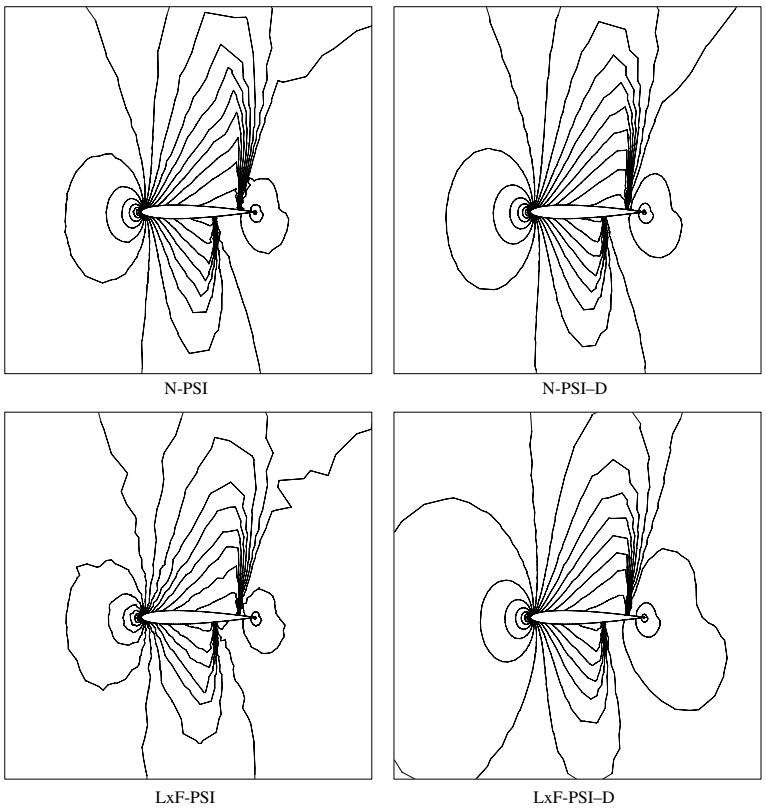

Fig. 22. NACA012 problem. Isolines of the density for the second order versions of the N and LxF schemes, without (left) and with (right) additional dissipation.

correction, these problems are cured. This result is obtained without sacrificing the quality of the discontinuities. This is also confirmed by the inspection of the other flow variables such as the pressure coefficient, the Mach number and the entropy deviation $(s - s_\infty)/s_\infty$.

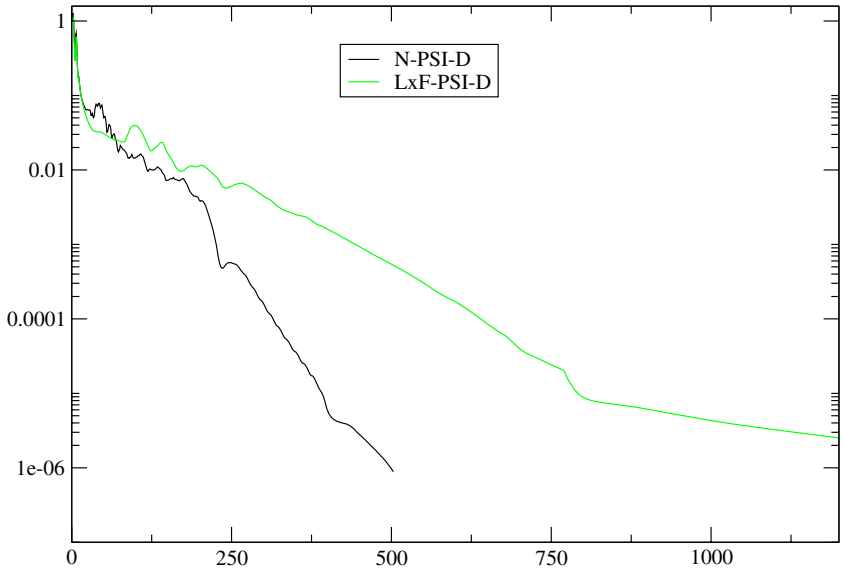

Fig. 23. Convergence history, CFL = 100 for the N-PSI-D scheme, CFL = 10 for the LxF-PSI-D scheme.

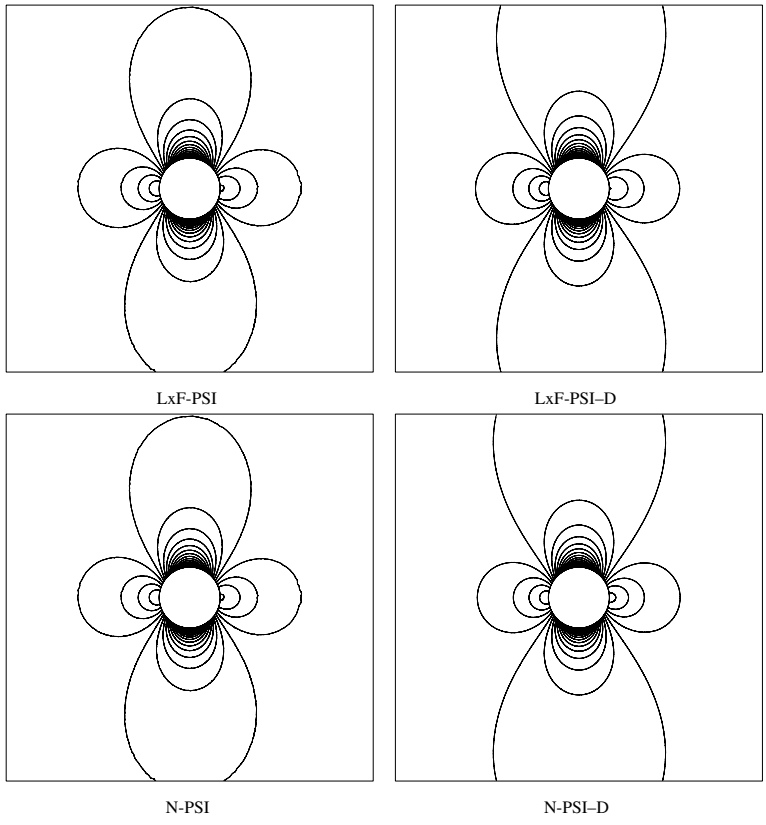

Fig. 24. C_p isolines for the sphere problem.

Last, the convergence history, to be compared with, is plotted in Fig. 23. The maximum CFL number for the N-PSI-D is set to 100 and only 10 for the LxF-PSI-D scheme. We have experimented that the second scheme is less robust in this case and the other we have run. However, we do not claim that our version of an implicit scheme is the best suited for the LxF scheme.

5.2.2. Subsonic flow

The inflow Mach number is set to $\mathscr{M}_\infty = 0.35$. The flow is subsonic everywhere. The pressure coefficient and the Mach number for the LxF-PSI and N-PSI schemes in Figs. 24 and 25 on the symmetric mesh plotted in Fig. 26(c). The results looks quite similar, but a close inspection of the isolines reveals some wiggles for the LxF-PSI scheme. The examination of the pressure coefficient contours for the LxF-PSI-D and N-PSI-D schemes, see Fig. 24, also show that they are more symmetric with respect to the y-axis compared to their non-dissipated counterparts. If we plot he pressure and Mach number on the sphere (not plotted), we see that the results of the LxF-PSI and N-PSI are not symmetrical. This is due to the poor convergence of the solution (the linear systems of the implicit phase are solved with a relative tolerance of 10^{-3}), see Fig. 27.

More interestingly, Fig. 26 show that the results of the LxF-PSI scheme is *very* dependent on the mesh quality. The mesh (a) is not symmetrical, and the results are extremely wiggly. These wiggles are completely cured for the LxF-PSI-D scheme (not shown). Note however that the mesh resolutions are similar.

The convergence histories are provided in Fig. 27. The maximum CFL here is set to 10. We see that the N-PSI-D has a better behavior than the LxF-PSI-D. The minimum residual is in between 10^{-5} and 10^{-6} and then stagnates. An examination of the residual isolines (not provided here) shows that this is likely due to the behavior of our implementation of the no-slip boundary conditions.

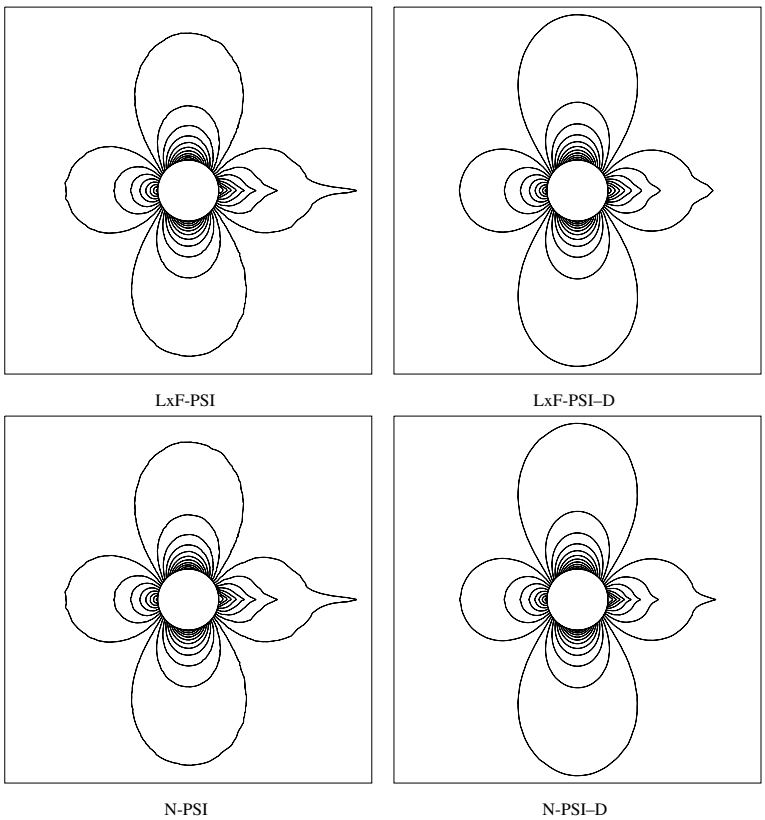

Fig. 25. Mach number isolines for the sphere problem.

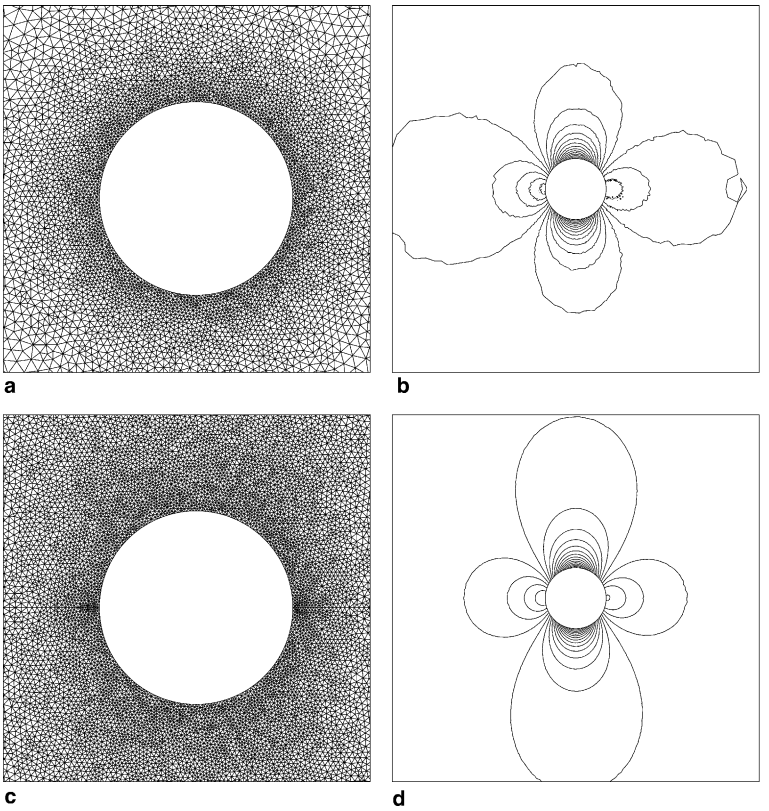

Fig. 26. Pressure coefficients for the LxF-PSI scheme on different meshes.

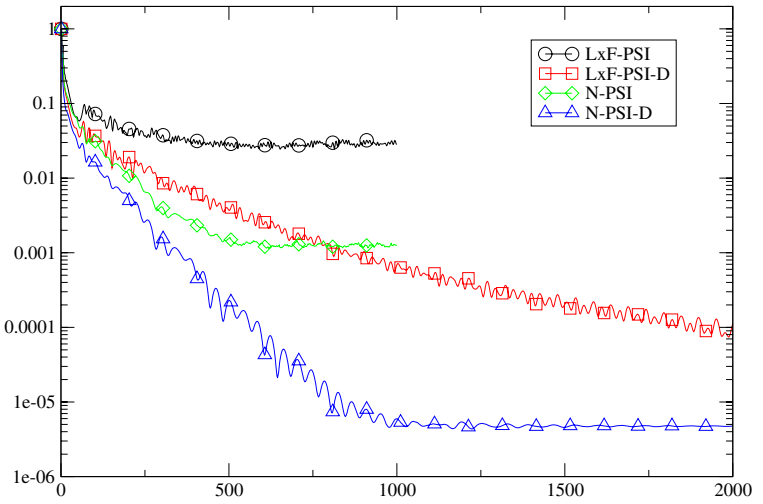

Fig. 27. Convergence history for the sphere problem.

5.2.3. Scramjet

Here, the inflow Mach number is $M = 3.6$. Because of the internal geometry, a very complex system of shock waves and slip lines occur, see Fig. 28. This make this example interesting since it permits to show the non-oscillatory behavior of the scheme in a rather complex configuration.

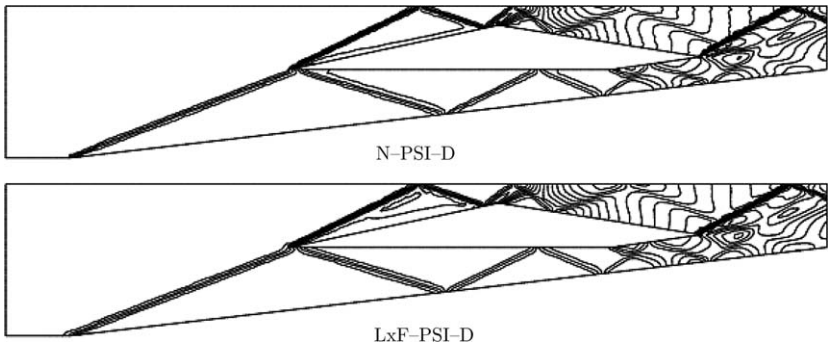

Fig. 28. Mach number isolines of the density for the N-PSI-D and the LxF-PSI-D schemes.

A zoom of the density isolines for each scheme is displayed in Figs. 29 and 30. This illustrates the perfect non-oscillatory behavior of the schemes even in rather complex configurations. In each case, the discontinuities are resolved within 2 cells.

Last, the convergence history is shown in Fig. 31 for two of the schemes. In this case, we had first to run the first order version of the schemes, and then the second order version with a maximum CFL of 10.

Other cases have been run, for example the flow over a sphere at Mach $\mathscr{M}_\infty = 8$ with good success. The results are not displayed here.

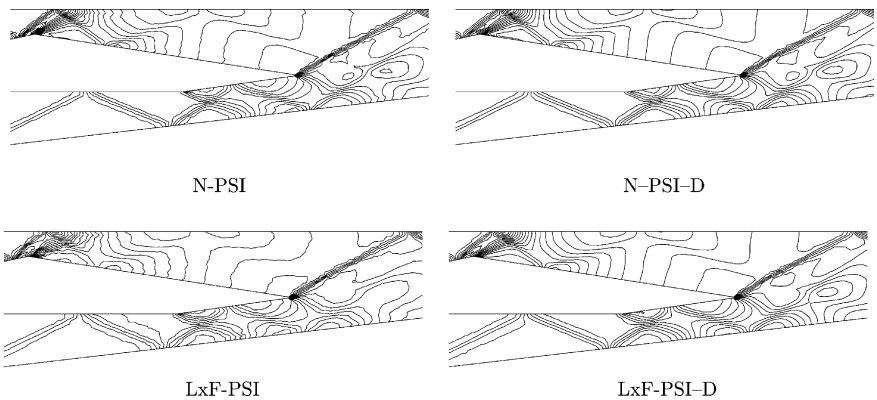

Fig. 29. Scramjet problem, zoom. Isolines of the density for the second order versions of the N, LxF schemes, without (left) and with (right) additional dissipation.

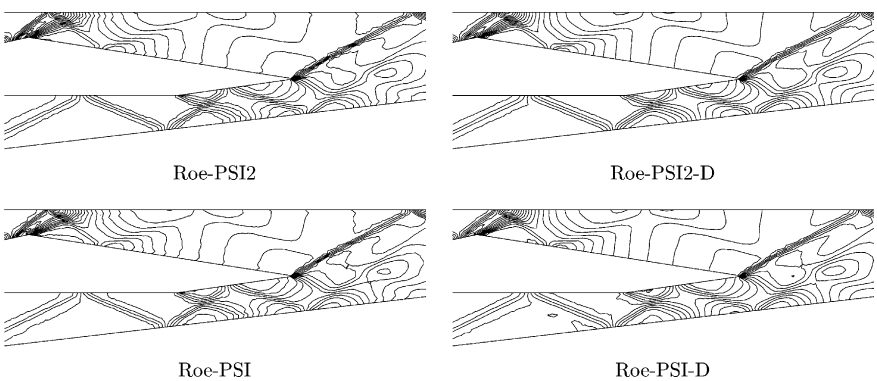

Fig. 30. Scramjet problem, zoom. Isolines of the density for the second order versions of the Roe and Roe2 schemes, without (left) and with (right) additional dissipation.

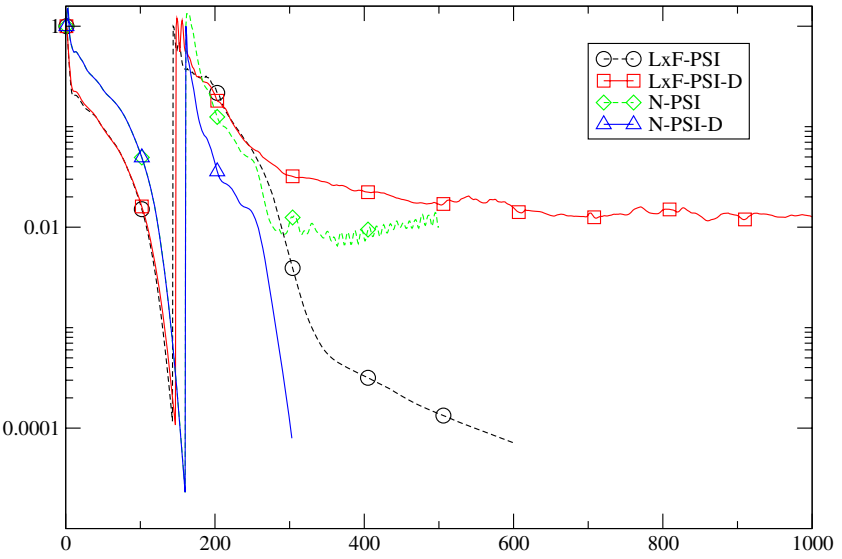

Fig. 31. Convergence history for the LxF and N-PSI type schemes for the scramjet problem.

6. Conclusions

This paper deals with the iterative convergence problem that is common to most monotonicity preserving Residual Distribution schemes for steady problems. This problem has been reported in several papers especially for systems, for example among others [5,4]. Since a good level of iterative convergence cannot be reached in general, the formal second order accuracy cannot be guarantied, since second order accuracy can only be obtained if and only if the residual Eq. (4) is solved exactly or with a tolerance of the order of the truncation error, provided the residuals Φ_i^T are defined by (8). A good convergence level is *essential*.

We first analyze the problem and connect it to the possible existence of spurious modes. Then we propose a solution, the price to pay is that the formal monotonicity of the scheme is lost. The technique is tested for several problems scalar and systems. Our results show that the fix we propose does not degrade the structure of discontinuities, a good convergence level is reached, in most case it also improves the quality of the solution in smooth parts (because of the better convergence of the iterative scheme).

This technique is extended with success to RD schemes for Cartesian meshes in [27]. Future work will consider the case of the unsteady problems following [3,28] and high order schemes too where similar difficulties are encountered.

Acknowledgements

It is my pleasure to thank Mohamed Mezine with whom I had many interesting and motivating discussions. This paper is certainly an attempt to answer some of them. Mario Ricchiuto also played an important role in this research because of his never-ending series of questions, remarks and enthousiasm. This research has been financed in part by contract 02 T 0313 of the French Ministry of Education and Research. Last the two referees are also thanked for their constructive remarks.

References

[1] Á. Csík, M. Ricchiuto, H. Deconinck, A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws, J. Comput. Phys. 179 (2) (2002) 286–312.
[2] R. Abgrall, P.L. Roe, Construction of very high order fluctuation scheme, J. Sci. Comput. 19 (1–3) (2003) 3–36.
[3] R. Abgrall, M. Mezine, Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems, J. Comput. Phys. 188 (1) (2003) 16–55.

[4] M. Ricchiuto, Construction and analysis of compact residual discretizations for conservation laws on unstructured meshes, PhD Thesis, Université Libre de Bruxelles, June 2005.
[5] R. Abgrall, Toward the ultimate conservative scheme: following the quest, J. Comput. Phys. 167 (2) (2001) 277–315.
[6] R. Struijs, H. Deconinck, P.L. Roe, Fluctuation splitting schemes for the 2D Euler equations, VKI LS 1991-01, Comput. Fluid Dynam., 1991.
[7] R. Abgrall, K. Mer, B. Nkonga, M. Hafez, A Lax–Wendroff type theorem for residual schemes, in: J.J Chattot (Ed.), Innovative Methods for Numerical Solutions of Partial Differential Equations, World Scientific, 2002, pp. 243–266.
[8] H. Deconinck, R. Struijs, G. Bourgeois, P.L. Roe, Compact advection schemes on unstructured meshes, VKI Lecture Series 1993-04, Comput. Fluid Dynam., 1993.
[9] H. Deconinck, R. Struijs, G. Bourgeois, P.L. Roe, Compact advection schemes on unstructured meshes, VKI Lecture Series 1993–04, Comput. Fluid Dynam., 1993.
[10] R. Struijs, H. Deconinck, P.L. Roe, Fluctuation splitting schemes for the 2D Euler equations, VKI LS 1991-01, 1991. Comput. Fluid Dynam.
[11] R. Abgrall, M. Mezine, Construction of second order accurate monotone and stable residual schemes for steady problems, J. Comput. Phys. 195 (2) (2004) 474–507.
[12] C. Johnson, Numerical Solution of Partial Differential Equations by the Finite Element Method, Cambridge University Press, Cambridge, 1987.
[13] C. Johnson, A. Szepessy, Convergence of the shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws, Math. Comput. 54 (1990) 107–129.
[14] T.J.R. Hughes, L.P. Franca, M. Mallet, A new finite element formulation for CFD: I. symmetric forms of the compressible Euler and Navier–Stokes equations and the second law of thermodynamics, Comput. Methods Appl. Mech. Engrg. 54 (1986) 223–234.
[15] T.J.R. Hughes, M. Mallet, A new finite element formulation for CFD: III. The generalized streamline operator for multidimensional advective–diffusive systems, Comput. Methods Appl. Mech. Engrg. 58 (1986) 305–328.
[16] R.-H. Ni, A multiple grid scheme for solving the Euler equations, AIAA J. 20 (1981) 1565–1571.
[17] K.W. Morton, E. Süli, Finite volume methods and their analysis, IMA J. Numer. Anal. 11 (1991) 241–260.
[18] P.I. Crumpton, J.A. MacKenzie, K.W. Morton, Cell vertex algorithms for the compressible Navier–Stokes equations, J. Comput. Phys. 109 (1993) 1–15.
[19] R. Abgrall, T.J. Barth, Weighted residual distribution schemes for conservation laws via adaptive quadrature, SIAM J. Sci. Comput. 24 (3) (2002) 732–769.
[20] E. van der Weide, H. Deconinck, Positive matrix distribution schemes for hyperbolic systems, in: Désidéri, Hirch, Le Tallec, Pandolfi, Périaux (Eds.), Computational Fluid Dynamics '96, Wiley, 1996, pp. 747–753.
[21] H. Deconinck, P.L. Roe, R. Struijs, A multidimensional generalisation of Roe's difference splitter for the Euler equations, Comput. Fluids 22 (2/3) (1993) 215–222.
[22] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes, J. Comput. Phys. 43 (1983).
[23] A. Dervieux, B. van Leer, J. Périaux, A. Rizzi (Eds.), Numerical Simulation of Compressible Euler Flows, Note on Numerical Fluid Mechanics, vol. 26, Vieweg, 1989.
[24] H. Paillère, Multidimensional upwind residual discretisation schemes for the Euler and Navier–Stokes equations on unstructured meshes, PhD Thesis, Université Libre de Bruxelles, 1995.
[25] M. Mezine, Conception de schémas distributifs pour l' aérodynamique stationnaire et instationnaire, PhD Thesis, Ecole Doctorale mathématique et Informatique, Université Bordeaux I, 2002.
[26] R. Abgrall, M. Mezine, Residual distribution schemes for steady problems, in: von Kàrman Institute Lecture Series, March 2003.
[27] F. Marpeau, R. Abgrall, Residual distribution schemes on quadrilateral meshes, J. Sci. Comput., 2005, in revision.
[28] M. Ricchiuto, Á. Csík, H. Deconinck, Residual distribution for general time dependant conservation laws, J. Comput. Phys. 209 (1) (2005) 249–289.