Numerische Mathematik manuscript No.
(will be inserted by the editor)

Stability Analysis of the Cell Centered Finite-Volume MUSCL Method on Unstructured Grids

F. Haider[1], **J-P. Croisille**[2], **B. Courbet**[1]

[1] ONERA, Département de simulation numérique des écoulements et aéroacoustique, 29 rue de la Division Leclerc, 92320 Châtillon FRANCE
e-mail: `florian.haider@onera.fr, bernard.courbet@onera.fr,`

[2] Laboratoire Mathématiques et Applications de Metz, UMR 7122 Université Paul Verlaine-Metz, Bât. A, Ile du Saulcy 57045 Metz FRANCE
e-mail: `jean-pierre.croisille@math.univ-metz.fr`

27 March 2009

Abstract The goal of this study is to apply the MUSCL scheme to the linear advection equation on general unstructured grids and to examine the eigenvalue stability of the resulting linear semi-discrete equation. Although this semi-discrete scheme is in general stable on cartesian grids, numerical calculations of spectra show that this can sometimes fail for generalizations of the MUSCL method to unstructured three-dimensional grids. This motivates our investigation of the influence of the slope reconstruction method and stencil on the eigenvalue stability of the MUSCL scheme. A theoretical stability analysis of the first order upwind scheme proves that this method is stable on arbitrary grids. In contrast, a general theoretical result is very difficult to obtain for the MUSCL scheme. We are able to identify a local property of the slope reconstruction that is strongly related to the appearance of unstable eigenmodes. This property allows to identify the reconstruction methods that are best suited for stable discretizations. The explicit numerical computation of spectra for a large number of two- and three-dimensional test cases confirms and completes the theoretical results.

Key words Finite volume scheme – stability analysis – hyperbolic equations – dynamical systems – method of lines – MUSCL method

Mathematics Subject Classification (2000): 65M12, 65M20, 76M12

1 Introduction

The finite volume MUSCL method to solve hyperbolic conservation laws was introduced by B. Van Leer in [53,54] thirty years ago. The main idea is to increase the accuracy of the first order finite volume scheme by a piecewise linear reconstruction that is used to evaluate upwinded fluxes at the cell interfaces. General references are [29,39,52,34,19,20], see also [15,5].

Practical applications for convection dominated flows in complex geometries have motivated many extensions of the MUSCL approach to general unstructured grids in 2 and 3 dimensions. In this context, one of the principal difficulties is to obtain high order accurate reconstructions that do not destroy the stability and robustness of the numerical scheme. The growing need for industrial applications of Large Eddy Simulations on unstructured grids makes this question more stringent than before. This topic is still an active field of research [2,31,36,51,11,16,35].

The purpose of this paper is to explore the relationship between the piecewise linear reconstruction and the stability of the MUSCL scheme. To isolate the influence of the former on the latter, the analysis proceeds in the following setting :

- We adopt the framework of the method of lines because it allows to study the stability of the MUSCL method regardless of the time stepping scheme. The application of the MUSCL scheme and the method of lines to a hyperbolic conservation law produces a dynamical system whose stability can be examined by the theory of dynamical systems.
- Slope limiters modify the piecewise linear reconstruction at the cell interfaces. In order to focus on the impact of the reconstruction step, we study the MUSCL scheme in the *absence of slope limiters*.
- We examine the specific case of the linear advection equation with constant velocity under periodic boundary conditions. Obviously, any implementation of the MUSCL scheme for this equation should result in a stable dynamical system. Furthermore, the linear advection equation is of primary importance to understand many properties of numerical schemes, such as accuracy, dispersion and stability.

In this context, the main interest of this paper is the *matrix stability* of the dynamical system that results from the application of the MUSCL scheme to the linear advection equation. As far as we know, this question is mathematically open on general unstructured grids. The goal of this paper is in particular to analyze the influence of the mesh type, the reconstruction method and the stencil size on the asymptotic stability of the dynamical system. A detailed discussion of the relationship between this kind of stability analysis and the classical *Von Neumann stability* analysis (Godunov-

Ryabenkii and Lax-Richtmeyer) can be found in [50,29,23,43]. Recall that on a periodic cartesian grid, the two points of view are equivalent [29].

It is now necessary to explain the interest of this specific notion of stability. We begin by the following observations about slope limiters and stability in order to relate the present work to the existing literature :

- Practical implementations of the MUSCL scheme need slope limiters to avoid oscillations near discontinuities. This step, which is necessary to restore the monotonicity near sharp fronts, was already present in the original work of Van Leer. Since the pioneering contribution by Harten [26], the quest for a general mathematical framework for the design of slope limiters became a very active field of research. ENO/WENO schemes go beyond a simple slope reconstruction by selecting the stencils leading to the least oscillating reconstruction, see for example [13, 1,33,48]. An important contribution in the framework of irregular general grids is due to Sonar who introduced an ENO slope reconstruction using radial basis functions to avoid the drawbacks of polynomial reconstruction, [49,32].
- Another important motivation for the design of slope limiters are conditions like L^p stability, entropy consistency, preservation of the positivity of physical quantities like mass density, see for example [10,42,7,8]. In certain cases, convergence towards a (the) weak solution of the conservation law can be proved. Error estimates are in general more difficult to obtain [17,12,41]. Concerning the finite volume scheme on unstructured grids, we refer to [3] for a rigorous proof of a maximum principle using a specific class of limiters. The result can be extended to the so called SSP Runge-Kutta methods, see [47,22].

These observations highlight the importance of slope limiters for the design of MUSCL schemes. However, the numerical dissipation introduced by slope limiters can be difficult to control in practical applications on unstructured grids. A particular example is the computation of a supersonic hot jet using LES turbulence modeling presented in [40]. On a purely tetrahedral grid, the numerical dissipation is too important for the jet to become turbulent and an excessive numerical dampening occurs. On a structured grid, however, the computation results in a more realistic unsteady solution. The same situation has been observed for the computation of a subsonic flow over a deep cavity on unstructured grids [9], albeit to a lesser extent. Note finally that the computation on the unstructured grid is less accurate than a similar computation carried out on structured grids by Larchevêque et al. [37].

In such a situation, it is natural to relax the monotonicity requirements and to use slope limiters that modify the piecewise linear reconstruction as slightly as possible. However, the computation of the subsonic flow over a deep cavity fails without the stronger slope limitation due to a lack of robustness. The comparison between structured and unstructured grids shows that the lack of robustness has to come from the piecewise linear reconstruction on unstructured grids, see [25,24].

This example shows that the piecewise linear reconstruction has a direct impact on the robustness and stability of the MUSCL scheme regardless of the slope limiters. We emphasize in particular that the MUSCL discretization of the linear advection equation using upwinded fluxes and a centered slope reconstruction on a uniform grid in 1 dimension is stable in the sense of matrix stability without any slope limitation, see [9]. This shows that there are situations where the piecewise linear reconstruction guarantees the stability of the scheme without the help of slope limiters.

The preceding observations suggest to proceed in the following way.

1. In a first step, investigate the influence of the piecewise linear reconstruction on the stability of the MUSCL scheme and identify the reconstruction methods that increase the robustness of the scheme. It seems in particular necessary to establish a local criterion that allows to compare different slope reconstructions regarding their impact on the global stability of the scheme, even if this criterion is only an approximate and qualitative one.
2. In a second step that is not covered by the present paper, it will be necessary to develop monotonicity requirements that add less numerical dissipation than those used for example in [9,40].

The outline is as follows. Subsection 2.1 presents the formulation of cell centered finite volume schemes for conservation laws under the method of lines. As there is no canonical way to reconstruct slopes on unstructured meshes, a general approach to this question is developed in Subsection 2.2. Subsection 3.1 outlines the MUSCL discretization of the advection equation and Subsection 3.2 presents the basic concepts of linear stability that are needed for our analysis. Subsection 3.3 describes the main theorem on the stability of the first order upwind finite volume scheme. Subsections 3.4, 3.5 and 3.6 present the analysis and the main results for the MUSCL scheme. These results lead to specific recommendations to enhance the stability properties of the MUSCL scheme concerning the choice of slope reconstruction on general unstructured grids, summarized in Subsection 3.7. Section 4 presents an extensive numerical study that has been performed with MAPLE. It completes the theoretical part by a range of interesting test cases covering several types of meshes in two and three dimensions.

Note finally that the present study is motivated by extensive numerical experiments in three-dimensional applications to internal flows and aerothermochemistry with the package CEDRE developed by ONERA. General references for CEDRE are [14,40,44,9,38,46].

The present paper was announced in [25]. See also [24] for more details.

2 Spatial Discretization on General Unstructured Meshes

2.1 General Formulation of Semi-discrete Finite-Volume Schemes

This section recalls the general setting of cell centered finite volume schemes on unstructured meshes in the context of the method of lines. Consider a conservation law

$$\partial_t u(\boldsymbol{x}, t) + \boldsymbol{\nabla} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) = 0 \tag{2.1}$$

where $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$ and $t \geq t_0$. Since our analysis does not address the specific influence of boundary conditions, they are assumed to be periodic. In the sequel, vectors in $\mathbb{R}^d$ will be written in bold letters like $\boldsymbol{c}$, matrices in capital letters and vectors in the space of semi-discrete solutions in fraktur like $\mathfrak{u}$.

The geometric notation is as follows. A general unstructured grid is a triangulation of Ω consisting of N general polyhedra,

$$\Omega = \bigcup_{\alpha=1}^{N} \mathcal{T}_\alpha \,.$$

The cell with number α is denoted $\mathcal{T}_\alpha$, with barycenter $\boldsymbol{x}_\alpha$ and d-volume $|\mathcal{T}_\alpha|$. The face $\mathcal{A}_{\alpha\beta}$, with barycenter $\boldsymbol{x}_{\alpha\beta}$, has a normal vector $\boldsymbol{a}_{\alpha\beta}$ oriented from cell $\mathcal{T}_\alpha$ to $\mathcal{T}_\beta$ and of length $\|\boldsymbol{a}_{\alpha\beta}\|$ equal to the surface $|\mathcal{A}_{\alpha\beta}|$. The oriented normal unit vector of the face $\mathcal{A}_{\alpha\beta}$ is $\boldsymbol{\nu}_{\alpha\beta}$.

Furthermore, the vectors $\boldsymbol{h}_{\alpha\beta}$, $\boldsymbol{k}_{\alpha\beta}$ are defined as

$$\begin{aligned} \boldsymbol{h}_{\alpha\beta} &= \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha \,; \text{ for all cells } \mathcal{T}_\alpha, \mathcal{T}_\beta \\ \boldsymbol{k}_{\alpha\beta} &= \boldsymbol{x}_{\alpha\beta} - \boldsymbol{x}_\alpha \,; \text{ for all adjacent cells } \mathcal{T}_\alpha, \mathcal{T}_\beta \,. \end{aligned}$$

The vector $\boldsymbol{j}_{\alpha\beta}$ is the orthogonal projection of $\boldsymbol{k}_{\alpha\beta}$ on $\boldsymbol{h}_{\alpha\beta}$ and $\boldsymbol{b}_{\alpha\beta}$ is defined by $\boldsymbol{k}_{\alpha\beta} = \boldsymbol{j}_{\alpha\beta} + \boldsymbol{b}_{\alpha\beta}$. The vector $\boldsymbol{j}_{\alpha\beta}$ is needed to define a slope reconstruction method tested in the numerical study of Section 4. Fig. 2.1 shows an example of the cell geometry.

Remark 2.1 (Curved faces) Note that the faces $\mathcal{A}_{\alpha\beta}$ need not be flat, i.e. the normal can vary from point to point. Such faces occur in three-dimensional irregular meshes whenever the vertices shared by two cells do not lie in a plane. The face is then spanned by the closed path formed by the line segments joining the vertices. In this case, even if the face itself is not unique, a normal vector can be uniquely defined as

$$\boldsymbol{a}_{\alpha\beta} = \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \, d\sigma$$

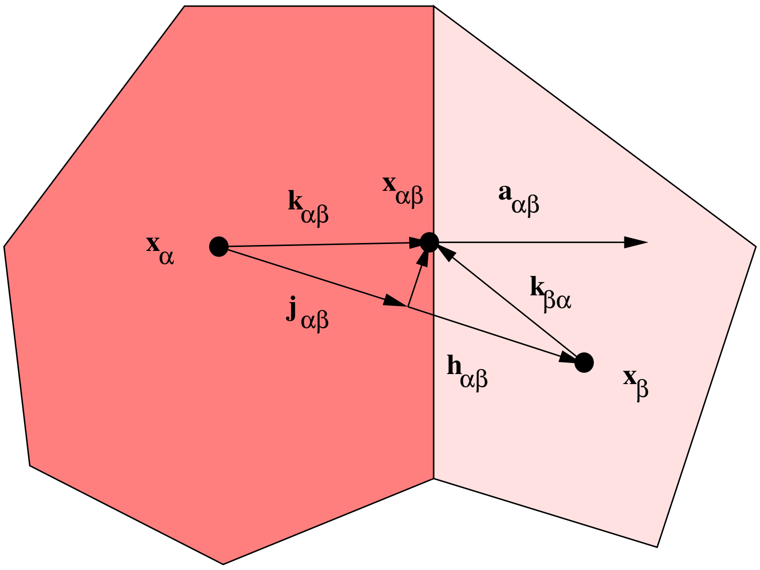

$\mathbf{x}_{\alpha}$
$\mathbf{k}_{\alpha\beta}$
$\mathbf{x}_{\alpha\beta}$
$\mathbf{a}_{\alpha\beta}$
$\mathbf{j}_{\alpha\beta}$
$\mathbf{k}_{\beta\alpha}$
$\mathbf{h}_{\alpha\beta}$
$\mathbf{x}_{\beta}$

Figure 2.1: Typical cell geometry of a general polyhedral grid

where $\boldsymbol{\nu}(\boldsymbol{x})$ is the unit normal in a point $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$. Green's Theorem guarantees that two faces delimited by the same closed path give the same $\boldsymbol{a}_{\alpha\beta}$.

Remark 2.2 As will be shown in Section 2.2, gradient reconstruction requires that at each cell $\mathcal{T}_\alpha$ the family $\{\boldsymbol{h}_{\alpha\beta}\}_{\beta\text{ adjacent to }\alpha}$ contain a set of d linearly independent vectors. This assumption is satisfied by all regular and irregular meshes used in practice.

The following convention simplifies the notation of sums over cells. Whenever two cells have no common interface, $\boldsymbol{a}_{\alpha\beta} = 0$ and $\boldsymbol{k}_{\alpha\beta} = 0$ and the face $\mathcal{A}_{\alpha\beta}$ is defined to be empty so that any surface integral over $\mathcal{A}_{\alpha\beta}$ is automatically zero. In addition, $\boldsymbol{a}_{\alpha\alpha}$, $\boldsymbol{k}_{\alpha\alpha}$ and $\boldsymbol{h}_{\alpha\alpha}$ are defined to be zero. This allows to drop the neighborhood in all sums and to write $\sum_\alpha$ instead of $\sum_{\beta\text{ adjacent to }\alpha}$. An example of this convention is the application of Green's Theorem to a constant function.

$$\begin{aligned} 0 &= \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla}(1)\ dx = \\ \sum_{\beta\text{ adjacent to }\alpha} \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x})\ d\sigma &= \sum_{\beta\text{ adjacent to }\alpha} \boldsymbol{a}_{\alpha\beta}\,. \end{aligned} \tag{2.2}$$

Now the sum in equation (2.2) can simply be written as

$$\sum_\beta \boldsymbol{a}_{\alpha\beta} = 0\,. \tag{2.3}$$

Note that identity (2.3) remains valid for cells with curved faces.

Integration of the conservation law (2.1) over any cell $\mathcal{T}_\alpha$ gives

$$\frac{d}{dt}\int_{\mathcal{T}_\alpha} u\left(\boldsymbol{x},t\right)\,dx = -\int_{\partial\mathcal{T}_\alpha} \boldsymbol{\nu}\cdot\boldsymbol{f}\left(u\left(\boldsymbol{x},t\right)\right)\,d\sigma \tag{2.4}$$

or equivalently

$$\frac{d\bar{u}_\alpha\left(t\right)}{dt} = -\frac{1}{\left|\mathcal{T}_\alpha\right|}\sum_\beta\int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}\left(u\left(\boldsymbol{x},t\right)\right)\,d\sigma \tag{2.5}$$

where the average of u over $\mathcal{T}_\alpha$ is

$$\bar{u}_\alpha\left(t\right) = \frac{1}{\left|\mathcal{T}_\alpha\right|}\int_{\mathcal{T}_\alpha} u\left(\boldsymbol{x},t\right)\,dx\,. \tag{2.6}$$

The simplest finite volume scheme consists in evolving the quantities $u_\alpha\left(t\right)$ approximating the exact averages $\bar{u}_\alpha\left(t\right)$ along the dynamical system

$$\frac{du_\alpha\left(t\right)}{dt} = -\frac{1}{\left|\mathcal{T}_\alpha\right|}\sum_\beta\int_{\mathcal{A}_{\alpha\beta}} \tilde{f}_{\alpha\beta}\left(u_\alpha\left(t\right), u_\beta\left(t\right)\right)\,d\sigma \tag{2.7}$$

where the numerical flux function $\tilde{f}_{\alpha\beta}\left(w_{\text{int}}, w_{\text{ext}}\right)$ depends on the two states w_{int} and w_{ext} on each side of the cell interface. We adopt the convention that for a face $\mathcal{A}_{\alpha\beta}$ oriented from cell $\mathcal{T}_\alpha$ to cell $\mathcal{T}_\beta$, w_{int} is the state on side α and w_{ext} the state on side β. The requirement

$$\tilde{f}_{\alpha\beta}\left(w_{\text{int}}, w_{\text{ext}}\right) = -\tilde{f}_{\beta\alpha}\left(w_{\text{ext}}, w_{\text{int}}\right)$$

guarantees the conservation of the total average

$$\frac{d}{dt}\left\{\sum_{\alpha=1}^{N}\left|\mathcal{T}_\alpha\right| u_\alpha\left(t\right)\right\} = 0\,. \tag{2.8}$$

The convergence of the scheme (2.7) usually requires the numerical flux to be consistent in the sense that

$$\tilde{f}_{\alpha\beta}\left(u,u\right) = \boldsymbol{f}\left(u\right)\cdot\boldsymbol{\nu}_{\alpha\beta}\quad\text{for all } u\in\mathbb{R}. \tag{2.9}$$

A proof of convergence for the time-discrete version of the scheme (2.7) in two dimensions can be found in [34, ch. 3.3]. A proof of strong convergence of the scheme (2.7) in the case of the linear advection equation is given in [17]. The convergence rate for this particular case has recently been improved in [41].

However, on a general unstructured mesh the scheme (2.7) has a local truncation error of order $O(1)$ as shown in [34, Lemma 3.2.8, page 161].

$$\frac{d\bar{u}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \tilde{f}_{\alpha\beta}\left(\bar{u}_\alpha(t), \bar{u}_\beta(t)\right)\, d\sigma + O(1) \tag{2.10}$$

To overcome this prohibitive lack of spatial accuracy in (2.10), the arguments $u_\alpha(t)$ and $u_\beta(t)$ in (2.7) are replaced by locally reconstructed values as follows. For a *fixed time* t_0 a function w is reconstructed from the time dependent cell averages $\mathfrak{u} = (u_1, \ldots, u_N)$ in a certain neighborhood of a given cell $\mathcal{T}_\alpha$. The dependence of w on the cell averages $\mathfrak{u}(t_0)$ is denoted by square brackets $w[\mathfrak{u}(t_0)]$ and the dependence on $\boldsymbol{x}$ by the usual parentheses $w[\mathfrak{u}(t_0)](\boldsymbol{x})$. The reconstruction process operates piecewise on each cell so that only the cell averages in a certain neighborhood of a cell $\mathcal{T}_\alpha$ determine the restriction of w on $\mathcal{T}_\alpha$. In the sequel, this restriction is denoted by w_α, such that w can be written as

$$w[\mathfrak{u}(t_0)](\boldsymbol{x}) = \sum_\alpha w_\alpha[\mathfrak{u}(t_0)](\boldsymbol{x})\,\chi_{\mathcal{T}_\alpha}(\boldsymbol{x}) \tag{2.11}$$

where $\chi_{\mathcal{T}_\alpha}(\boldsymbol{x})$ is the characteristic function of the cell $\mathcal{T}_\alpha$.

To produce a useful scheme, the reconstruction must satisfy the two following requirements.

- Conservation
$$\frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} w[\mathfrak{u}(t_0)](\boldsymbol{x})\, dx = u_\alpha\,;\ 1 \leq \alpha \leq N \tag{2.12}$$
- Accuracy
$$|w[\mathfrak{u}(t_0)](\boldsymbol{x}) - u(\boldsymbol{x}, t_0)| \leq O(h^p) \tag{2.13}$$

The relation (2.13) must hold uniformly in $\boldsymbol{x} \in \Omega$ for all sufficiently smooth functions u with cell averages $\mathfrak{u}$. In (2.13), h is the maximum diameter of the mesh cells. The integer p is called *the order of the reconstruction*. The scheme deduced from (2.7) using the reconstruction is now

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \tilde{f}_{\alpha\beta}\left(w_\alpha[\mathfrak{u}(t)](\boldsymbol{x}), w_\beta[\mathfrak{u}(t)](\boldsymbol{x})\right)\, d\sigma \tag{2.14}$$

which can be shown to be of consistency $O\left(h^{p-1}\right)$ if the numerical flux is Lipschitz-continuous in both arguments. The final step is to approximate the integral in (2.14) by an appropriate quadrature formula which gives the dynamical system

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \sum_q \omega_q\, \tilde{f}_{\alpha\beta}\left(w_\alpha[\mathfrak{u}(t)](\boldsymbol{x}_{\alpha\beta;q}), w_\beta[\mathfrak{u}(t)](\boldsymbol{x}_{\alpha\beta;q})\right)\,. \tag{2.15}$$

In equation (2.15), the $\boldsymbol{x}_{\alpha\beta;q}$ are the quadrature points on $\mathcal{A}_{\alpha\beta}$ and the ω_q are the quadrature weights. If the expression under the integral on the right hand side of (2.14) is a polynomial of degree p in $\boldsymbol{x}$ and if the quadrature formula integrates exactly such polynomials, this step does not introduce any new discretization errors.

2.2 Analysis of Gradient Reconstruction on General Unstructured Meshes

In practical applications, the piecewise linear reconstruction of the MUSCL scheme is known to provide a radical improvement in spatial accuracy when compared with the first order scheme (2.7). Reconstructing a slope on a cartesian mesh is quite easy using directional splitting. However, on general unstructured meshes there is no canonical method.

A piecewise linear reconstructed function w is defined as

$$w_\alpha \left[\mathfrak{u}\left(t\right)\right]\left(\boldsymbol{x}\right) = u_\alpha\left(t\right) + \boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\left(t\right)\right]\cdot\left(\boldsymbol{x}-\boldsymbol{x}_\alpha\right) \text{ for } \boldsymbol{x}\in\mathcal{T}_\alpha \tag{2.16}$$

where the cell gradient $\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]$ is obtained from the cell averages $\mathfrak{u} = \left(u_1,\ldots,u_N\right)$ in a neighborhood of the cell. Equation (2.16) shows that the problem of piecewise linear reconstruction is equivalent to the problem of slope reconstruction from cell averages.

In order for (2.16) to fulfill the accuracy requirement according to (2.13) with $p=2$, we define a consistent slope reconstruction as

Definition 2.3 (Consistent Slope Reconstruction) *A piecewise linear reconstruction is called* consistent *on a neighborhood of a cell $\mathcal{T}_\alpha$ if for any affine function u on this neighborhood we have*

$$w_\alpha\left[\mathfrak{u}\right]\left(\boldsymbol{x}\right) = u\left(\boldsymbol{x}\right)$$

where $\mathfrak{u}$ is the vector of cell averages of the function u.

The reconstruction stencil of the gradient $\boldsymbol{\sigma}_\alpha$ in cell $\mathcal{T}_\alpha$ is the set of cells $\mathcal{T}_\beta$ such that $\boldsymbol{\sigma}_\alpha$ depends on u_β. The *first neighborhood* of a cell $\mathcal{T}_\alpha$ is the set of all cells sharing a common face with $\mathcal{T}_\alpha$. The *second neighborhood* of a cell $\mathcal{T}_\alpha$ is defined as the union of the first neighborhoods of the first neighbors of $\mathcal{T}_\alpha$, excluding $\mathcal{T}_\alpha$ itself.

Taking the gradient and the cell average of a function are both linear operations. Consistent gradient reconstruction is the inverse operation of the cell average on the space of polynomials of degree one. This justifies to focus on gradient reconstruction methods with linear dependence

$$\mathfrak{u}\mapsto\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{s}_{\alpha\beta}\left(u_\beta - u_\alpha\right) \ . \tag{2.17}$$

The vectors $\boldsymbol{s}_{\alpha\beta}$ in (2.17) are parameter vectors in cell $\mathcal{T}_\alpha$. Again $\boldsymbol{s}_{\alpha\beta} \triangleq 0$ if cell $\mathcal{T}_\beta$ is not in the reconstruction stencil of cell $\mathcal{T}_\alpha$.

If the underlying function u is linear with gradient $\boldsymbol{\sigma} \in \mathbb{R}^d$ then $u_\beta = u_\alpha + \boldsymbol{\sigma} \cdot \boldsymbol{h}_{\alpha\beta}$. The condition of first order consistency can therefore be written as

$$\boldsymbol{\sigma} = \sum_\beta \boldsymbol{s}_{\alpha\beta}\ (\boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{\sigma}) \quad \text{for all}\, \boldsymbol{\sigma} \in \mathbb{R}^d\,. \tag{2.18}$$

This is equivalent to the matrix identity

$$\sum_\beta \boldsymbol{s}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{I}_{d\times d}\,. \tag{2.19}$$

Let m_α be the number of cells in the reconstruction stencil of cell α and $\mathbb{W}_\alpha \triangleq \{\beta_1, \beta_2, \ldots, \beta_{m_\alpha}\}$ the cell indexes in that stencil. On cell $\mathcal{T}_\alpha$, the unknown vectors $\boldsymbol{s}_{\alpha\beta}$ form the columns of a $d \times m_\alpha$ matrix S_α. Similarly, the metric vectors $\boldsymbol{h}_{\alpha\beta}$ form the rows of the $m_\alpha \times d$ matrix H_α

$$H_\alpha^t = [\boldsymbol{h}_{\alpha\beta_1}, \boldsymbol{h}_{\alpha\beta_2}, \ldots, \boldsymbol{h}_{\alpha\beta_m}] \tag{2.20}$$

$$S_\alpha = [\boldsymbol{s}_{\alpha\beta_1}, \boldsymbol{s}_{\alpha\beta_2}, \ldots, \boldsymbol{s}_{\alpha\beta_m}]\,. \tag{2.21}$$

Equation (2.19) can be written as the matrix equation with unknown S_α

$$S_\alpha H_\alpha = \boldsymbol{I}_{d\times d}\,. \tag{2.22}$$

On the assumption that rank $(H_\alpha) = d$, (see Remark 2.2 above), the general solution is expressed as

$$S_\alpha = \tilde{S}_\alpha + \Lambda_\alpha B_\alpha \tag{2.23}$$

where $\tilde{S}_\alpha$ is a particular solution and B_α a maximal rank solution of the homogeneous equation ($\boldsymbol{O}_{(m_\alpha - d)\times d}$ is the zero matrix)

$$B_\alpha H_\alpha = \boldsymbol{O}_{(m_\alpha - d)\times d}\,. \tag{2.24}$$

Λ_α is an arbitrary matrix of size $d \times (m_\alpha - d)$ representing the degrees of freedom of the consistent reconstruction in cell $\mathcal{T}_\alpha$. Different choices of Λ_α lead to different consistent reconstruction methods.

The popular least squares reconstruction of the cell gradient is specified in the following

Proposition 2.4 (Least-Squares Reconstruction) *Let the matrix H_α have rank d and let $\boldsymbol{\sigma}_\alpha \in \mathbb{R}^d$ be the solution of the least squares problem*

$$\min_{\boldsymbol{\sigma}\in\mathbb{R}^d} \left\{ \sum_{\beta\in\mathbb{W}_\alpha} (u_\beta - u_\alpha - \boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{\sigma})^2 \right\}\,. \tag{2.25}$$

Then $\boldsymbol{\sigma}_\alpha$ is unique and given by coefficients $\boldsymbol{s}_{\alpha\beta}$ that are the columns of the minimum Frobenius norm solution to equation (2.22).

Proof : The solution $\boldsymbol{\sigma}_\alpha$ of Problem (2.25) satisfies the equation

$$\sum_{\beta\in\mathbb{W}_\alpha} \boldsymbol{h}_{\alpha\beta}\left(\boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{\sigma}\right) = \sum_{\beta\in\mathbb{W}_\alpha} \boldsymbol{h}_{\alpha\beta}\left(u_\beta - u_\alpha\right)\ . \tag{2.26}$$

Let $\boldsymbol{\delta u} = \left(u_{\beta_1} - u_\alpha, \dots, u_{\beta_{m_\alpha}} - u_\alpha\right)$. Then the unique solution of (2.26) is

$$\boldsymbol{\sigma} = \tilde{S}_\alpha \boldsymbol{\delta u} = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t \boldsymbol{\delta u}\,. \tag{2.27}$$

The matrix $\tilde{S}_\alpha$ is the pseudo-inverse of H_α that is known to minimize the Frobenius norm $\|S_\alpha\|_F = \sqrt{\operatorname{tr}\left(S_\alpha^t S_\alpha\right)}$ among the solutions of (2.22), see [21, Chap.5.5.4, pp 257 sqq.].

□

3 Stability Analysis of the Semi-discrete MUSCL scheme

This section presents the stability analysis of the semi-discrete MUSCL scheme for the linear advection equation with constant velocity. In Subsection 3.1, we derive the semi-discrete equations. Subsection 3.2 introduces preliminary concepts of linear stability that are needed later. A special emphasis is put on a variant of the classical Lyapunov Theorem. Subsection 3.3 contains a general stability result for the upwinded first order finite volume scheme on arbitrary meshes. Finally, in Subsections 3.4, 3.5 and 3.6 we discuss the main results of this paper.

3.1 Construction of the Semi-discrete Equations

The linear advection equation with constant velocity $\boldsymbol{c}\in\mathbb{R}^d$ is

$$\partial_t u\left(\boldsymbol{x},t\right) + \boldsymbol{c}\cdot\boldsymbol{\nabla} u\left(\boldsymbol{x},t\right) = 0\,,\ \left(\boldsymbol{x},t\right)\in\mathbb{R}^d\times\mathbb{R}_+\,. \tag{3.1}$$

Its conservative flux is linear in u

$$\boldsymbol{f}\left(u\right) = \boldsymbol{c}u. \tag{3.2}$$

The discretization of (3.1) follows the lines of Subsection 2.1. The usual upwinded numerical flux used in (2.14) is

$$\tilde{f}_{\alpha\beta}\left(w_{\text{int}}, w_{\text{ext}}\right) = \mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\frac{w_{\text{int}} + w_{\text{ext}}}{2} + \left|\mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\right|\frac{w_{\text{int}} - w_{\text{ext}}}{2}\ . \tag{3.3}$$

The use of piecewise constant and linear reconstruction in (2.14) leads to the following two dynamical systems :

- The first order finite volume scheme

$$\frac{du_\alpha(t)}{dt} = \sum_\beta \tilde{J}_{\alpha\beta} u_\beta(t) \; ; \; 1 \le \alpha \le N. \tag{3.4}$$

- The MUSCL finite volume scheme

$$\frac{du_\alpha(t)}{dt} = \sum_\beta J_{\alpha\beta} u_\beta(t) \; ; \; 1 \le \alpha \le N. \tag{3.5}$$

The spatial discretization operator $\widetilde{J}$ in (3.4) is given by

$$\tilde{J}_{\alpha\beta} = -\frac{1}{|\mathcal{T}_\alpha|} \sum_\gamma (\mathbf{c} \cdot \boldsymbol{a}_{\alpha\gamma})_+ \delta_{\alpha\beta} - \frac{1}{|\mathcal{T}_\alpha|} (\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \tag{3.6}$$

where $\delta_{\alpha\beta}$ is the Kronecker symbol. The operator J of the MUSCL scheme (3.5) is

$$\begin{aligned} J_{\alpha\beta} = -\frac{1}{|\mathcal{T}_\alpha|} \Bigg\{ & \sum_\gamma (\mathbf{c} \cdot \boldsymbol{a}_{\alpha\gamma})_+ \delta_{\alpha\beta} + (\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \\ & + \sum_\gamma (\boldsymbol{a}_{\alpha\gamma} \cdot \boldsymbol{c})_+ \boldsymbol{k}_{\alpha\gamma} \cdot \boldsymbol{s}_{\alpha\beta} - \sum_\gamma (\boldsymbol{a}_{\alpha\gamma} \cdot \boldsymbol{c})_+ \boldsymbol{k}_{\alpha\gamma} \cdot \boldsymbol{s}_\alpha \, \delta_{\alpha\beta} \\ & - \sum_\gamma (\boldsymbol{a}_{\gamma\alpha} \cdot \boldsymbol{c})_+ \boldsymbol{k}_{\gamma\alpha} \cdot \boldsymbol{s}_{\gamma\beta} + (\boldsymbol{a}_{\beta\alpha} \cdot \boldsymbol{c})_+ \boldsymbol{k}_{\beta\alpha} \cdot \boldsymbol{s}_\beta \Bigg\} \; . \end{aligned} \tag{3.7}$$

where $\boldsymbol{s}_\alpha \triangleq \sum_\beta \boldsymbol{s}_{\alpha\beta}$. The right hand side of (3.7) is obtained by inserting in the dynamical system (2.14) the reconstructed functions

$$\begin{aligned} w_\alpha &= u_\alpha + (\boldsymbol{x} - \boldsymbol{x}_\alpha) \cdot \sum_\gamma \boldsymbol{s}_{\alpha\gamma} (u_\gamma - u_\alpha) \\ w_\beta &= u_\beta + (\boldsymbol{x} - \boldsymbol{x}_\beta) \cdot \sum_\gamma \boldsymbol{s}_{\beta\gamma} (u_\gamma - u_\beta) \end{aligned} \tag{3.8}$$

and the numerical flux (3.3) in the following way.

$$\begin{aligned} |\mathcal{T}_\alpha| \frac{du_\alpha}{dt} = & \\ & - \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_+ \left\{ u_\alpha + (\boldsymbol{x} - \boldsymbol{x}_\alpha) \cdot \sum_\gamma \boldsymbol{s}_{\alpha\gamma} (u_\gamma - u_\alpha) \right\} d\sigma \\ & - \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_- \left\{ u_\beta + (\boldsymbol{x} - \boldsymbol{x}_\beta) \cdot \sum_\gamma \boldsymbol{s}_{\beta\gamma} (u_\gamma - u_\beta) \right\} d\sigma \; . \end{aligned} \tag{3.9}$$

When the face $\mathcal{A}_{\alpha\beta}$ is flat, i.e. $\boldsymbol{\nu}(x) = \boldsymbol{\nu}_{\alpha\beta} =$ const., the operator (3.7) follows from (3.9) because of the relation

$$\int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_{\pm} (\boldsymbol{x} - \boldsymbol{x}_{\alpha})\, d\sigma = (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_{\pm} \boldsymbol{k}_{\alpha\beta}\,. \tag{3.10}$$

If the face is curved, (see Remark 2.1 above), (3.7) is an approximation of (2.14). The first order operator (3.6) is always exact since

$$\int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_{\pm}\, d\sigma = (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_{\pm} \tag{3.11}$$

holds both for flat and curved faces.

3.2 Preliminary Concepts of Linear Stability

Recall that the linear advection equation preserves the norm of any initial condition u_0. Consider for example periodic boundary conditions in $\Omega = [0,1]^d$. The form of the solution $u(\boldsymbol{x},t) = u_0(\boldsymbol{x} - \boldsymbol{c}t)$ implies the conservation of all norms $L_p(\Omega)$, $1 \le p \le \infty$,

$$\|u(.,t)\|_{L_p(\Omega)} = \|u_0\|_{L_p(\Omega)}\,. \tag{3.12}$$

Let us examine to which extent property (3.12) can be preserved by the schemes (3.6) and (3.7).

As mentioned at the end of Section 1, the main interest of this work is the matrix stability analysis in the usual sense of (3.6) and (3.7), see [29, vol. 1]. In general, this kind of analysis leads to stability conditions that are different from those given by the Lax stability analysis. However, it is well known that the two kinds of stability conditions are the same for periodic problems on cartesian grids, see [29, vol. 1].

Let us start by recalling the following result for a general linear homogeneous autonomous system

$$\frac{d\mathrm{u}(t)}{dt} = J\mathrm{u}(t)\,,\ \mathrm{u}(0) = \mathrm{u}_0\,,\ \mathrm{u}(t) \in \mathbb{C}^N\,,\ J \in \mathbb{M}_N(\mathbb{C}) \tag{3.13}$$

whose solution is $\mathrm{u}(t) = \exp(tJ)\,\mathrm{u}_0$, see [18, Lm. 3.20 on p.95 and Th. 3.23 on p. 97].

Proposition 3.1 (Stability of Linear Systems in Finite Dimension) *The system (3.13) is stable in the sense that*

$$\mathrm{C} = \sup_{t \ge 0} \|\exp(tJ)\| < \infty \tag{3.14}$$

if and only if all eigenvalues λ of J satisfy :

(i) $\Re(\lambda) \leq 0$ *where* $\Re(\lambda)$ *is the real part of* λ.

(ii) if $\Re(\lambda) = 0$ *then the Jordan index* $\imath(\lambda) = 1$ *where* $\imath(\lambda)$ *is the maximal dimension of the Jordan blocks of* J *containing* λ.

In particular, the solution of a stable system satisfies

$$\sup_{t \geq 0} \|\mathfrak{u}(t)\| \leq \mathrm{C} \|\mathfrak{u}_0\| \ . \tag{3.15}$$

Property (3.15) has to be satisfied as a minimal requirement by the semi-discrete finite volume operators (3.6) and (3.7) independently of the advection velocity $\boldsymbol{c} \in \mathbb{R}^d$.

Definition 3.2 (Stable finite volume operator) *The spatial discretization operators* $\widetilde{J}$ *in (3.6) and* J *in (3.7) are called* stable *if all their eigenvalues satisfy properties (i)-(ii) of Proposition 3.1 for* all *advection velocities* $\boldsymbol{c} \in \mathbb{R}^d$.

For the convenience of the reader, the following theorem collects several elementary facts about the location of the spectrum of a matrix $A \in \mathbb{M}_N(\mathbb{C})$. Proofs can be found in [55, Theorem 1.1., p. 4, Theorem 3.7., p. 79], [27, Theorem 6.6.1, p. 344] and [28, Property 1.2.6, p. 10].

Theorem 3.3 (Geršgorin Disks, Field of Values)

(i) The spectrum of A *is contained in the union of all Geršgorin disks*

$$\sigma(A) \subseteq \bigcup_{i=1}^{N} \mathit{\Gamma}_i(A)$$

where the i*-th Geršgorin* disk *of* A *is defined by*

$$\mathit{\Gamma}_i(A) = \left\{ z \in \mathbb{C} \ : \ |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\} \ .$$

(ii) the spectrum of A *lies within the field of values of* A

$$\sigma(A) \subseteq \mathfrak{F}(A) \ .$$

The field of values $\mathfrak{F}(A)$ *is defined by*

$$\mathfrak{F}(A) = \left\{ (\boldsymbol{z}, A\boldsymbol{z}) \ : \ \boldsymbol{z} \in \mathbb{C}^N \ , \ (\boldsymbol{z}, \boldsymbol{z}) = 1 \right\}$$

where $(.,.)$ *denotes the Hermitian product on* $\mathbb{C}^N$.

□

Theorem 3.3 turned out to be insufficient to locate accurately the spectrum of the operator (3.7). Even on a regular mesh in one dimension, each Geršgorin disk of the MUSCL operator J in (3.7) intersects the complex right half plane although the spectrum is contained in the left half plane. We have examined other eigenvalue inclusion sets presented in [55], but unfortunately the result remains the same.

This observation led to a different approach based on the classical Lyapunov Theorem, [28, Theorem 2.2.1 on page 96]. Actually, we need the following variant.

Theorem 3.4 (Extended Lyapunov Theorem) *Let* $J \in \mathbb{M}_N(\mathbb{C})$. *The following properties are equivalent :*

(i) J satisfies the conditions of Proposition 3.1, that is all eigenvalues λ of J have $\Re(\lambda) \leq 0$ and if $\Re(\lambda) = 0$ then the Jordan index $\imath(\lambda) = 1$.

(ii) There exists a positive definite matrix G such that the matrix $Q = GJ + J^*G$ is negative semidefinite.

Proof : First observe that properties (i) and (ii) are true for a matrix J if and only if they are true for any matrix similar to J. This is clear for property (i). Now assume that J has property (ii) and let $\hat{J} = S^{-1}JS$ be a matrix similar to J. Then

$$S^*GS\,S^{-1}JS + S^*J^*\left(S^*\right)^{-1}\,S^*GS = S^*QS$$

Thus property (ii) is true for $\hat{J}$ with matrices $\hat{Q} = S^*QS$ and $\hat{G} = S^*GS$ because $\hat{Q}$ being congruent to Q is negative semidefinite and $\hat{G}$ being congruent to G is positive definite.

Proof of (i) $\Rightarrow$(ii) : Let J have property (i). According to the preliminary observation one can assume J to be in Jordan normal form $\hat{J}$. For all Jordan blocks associated with eigenvalues λ with $\Re(\lambda) < 0$ one can suppose that the supra diagonal elements are equal to an $\varepsilon > 0$ instead of 1, see [27, Corollary 3.1.13, page 128]. Choose an ε satisfying

$$\varepsilon < \min\left\{|\Re(\lambda)|\ ,\ \lambda \in \sigma(J)\ ,\ \Re(\lambda) < 0\right\}\ . \tag{3.16}$$

All Jordan blocks corresponding to eigenvalues λ with $\Re(\lambda) = 0$ are diagonal because $\imath(\lambda) = 1$. The matrix $\hat{Q} = \hat{J} + \hat{J}^*$ is Hermitian and in block diagonal form. Relation (3.16) ensures that the blocks of $\hat{Q}$ associated with λ such that $\Re(\lambda) < 0$ are strictly diagonally dominant. The blocks of $\hat{Q}$ associated with λ such that $\Re(\lambda) = 0$ are zero. The matrix $\hat{Q}$ is therefore negative semidefinite which proves that the Jordan normal form $\hat{J}$ has property (ii) with $\hat{G}$ being the identity matrix and $\hat{Q} = \hat{J} + \hat{J}^*$. Since the matrix J is similar to $\hat{J}$, it has property (ii).

Proof of (ii) ⇒(i) : Assume that J has property (ii). Multiplying the relation $GJ + J^{*}G = Q$ by $G^{-\frac{1}{2}}$ on both sides gives

$$G^{\frac{1}{2}}JG^{-\frac{1}{2}} + G^{-\frac{1}{2}}J^{*}G^{\frac{1}{2}} = G^{\frac{1}{2}}JG^{-\frac{1}{2}} + \left(G^{\frac{1}{2}}JG^{-\frac{1}{2}}\right)^{*} = G^{-\frac{1}{2}}QG^{-\frac{1}{2}} \,.$$

This means that the Hermitian part of the matrix $G^{\frac{1}{2}}JG^{-\frac{1}{2}}$ is negative semidefinite. Thus, all eigenvalues of this matrix have non positive real parts. The same holds true for the eigenvalues of J because it is similar to $G^{\frac{1}{2}}JG^{-\frac{1}{2}}$.

It remains to show that any purely imaginary eigenvalue λ of J has a Jordan index $\imath\,(\lambda) = 1$. According to the observation at the beginning of the proof one can suppose J to be in Jordan normal form. Assume that there is an eigenvalue $\lambda = i\mu\,,\ \mu \in \mathbb{R}$, associated with the l-th Jordan block $J^{(l)}$ of J. Suppose further that $J^{(l)}$ is not diagonalizable. Because of the block structure of the Jordan normal form of J the relation $GJ+J^{*}G = Q$ can be written block by block. Let $G^{(l)}$ be the diagonal block of G corresponding to $J^{(l)}$. The block $G^{(l)}$ must be positive definite as a diagonal block of G. Then $G^{(l)}J^{(l)} + J^{(l)*}G^{(l)} = Q^{(l)}$ where $Q^{(l)}$ is the corresponding diagonal block of Q. $Q^{(l)}$ must be negative semidefinite as a diagonal block of Q. The relation $G^{(l)}J^{(l)} + J^{(l)*}G^{(l)} = Q^{(l)}$ can be written in explicit form as

$$\begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \\ \vdots & & \ddots \end{pmatrix}\begin{pmatrix} i\mu & 1 & 0 & \cdots \\ 0 & i\mu & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} +$$

$$\begin{pmatrix} -i\mu & 0 & \cdots \\ 1 & -i\mu & \ddots \\ 0 & 1 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}\begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \\ \vdots & & \ddots \end{pmatrix} = \begin{pmatrix} 0 & g_{11} & \cdots \\ g_{11} & g_{12}+g_{21} & \\ \vdots & & \ddots \end{pmatrix}$$

In particular, $Q^{(l)}_{11} = 0$ and $Q^{(l)}_{12} = g_{11}$. But $g_{11} > 0$ since G is positive definite. However, a matrix $Q^{(l)}$ with $Q^{(l)}_{11} = 0$ and $Q^{(l)}_{12} > 0$ cannot be semidefinite. This contradiction shows that the Jordan block associated with λ must be diagonal.

□

Combining Proposition 3.1 and Theorem 3.4 yields the following

Corollary 3.5 *Consider the initial value problem*

$$\frac{d\mathfrak{u}\,(t)}{dt} = J\mathfrak{u}\,(t)\ ,\ \mathfrak{u}\,(0) = \mathfrak{u}_0\,,\ \mathfrak{u}\,(t) \in \mathbb{C}^{N}\ ,\ J \in \mathbb{M}_N\,(\mathbb{C})\ . \qquad (3.17)$$

Then there is a constant C *such that*

$$\|\mathfrak{u}\,(t)\| \leq \mathrm{C}\,\|\mathfrak{u}_0\| \text{ for all } t \geq 0$$

if and only if there exists a positive definite matrix G *such that the matrix* $Q = GJ + J^*G$ *is negative semidefinite.*

3.3 Stability Analysis of the First Order Finite-Volume Scheme

This section deals with the stability analysis of the dynamical system (3.4)

$$\frac{du_\alpha\,(t)}{dt} = \sum_\beta \widetilde{J}_{\alpha\beta} u_\beta\,(t)\ ;\ 1 \leq \alpha \leq N$$

defined by the first order finite volume operator (3.6)

$$\widetilde{J}_{\alpha\beta} = -\,|\mathcal{T}_\alpha|^{-1} \left\{ \sum_\gamma \left(\mathbf{c} \cdot \boldsymbol{a}_{\alpha\gamma}\right)_+ \delta_{\alpha\beta} + \left(\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta}\right)_- \right\}\,.$$

Although the matrix of (3.6) is real valued, the operator (3.6) has in general complex eigenvalues. For this reason it is preferable to work with complex vectors $\mathfrak{u} \in \mathbb{C}^N$. The complex conjugate of $\mathfrak{u} \in \mathbb{C}^N$ is denoted by $\mathfrak{u}^*$.

Observe first that all eigenvalues of $\widetilde{J}$ lie in the left closed half plane of the complex numbers. Indeed, for all cells $\mathcal{T}_\alpha$, relation (2.3) implies the existence of a face $\mathcal{A}_{\alpha\beta}$ such that $\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta} > 0$. Therefore the diagonal elements of (3.6) satisfy

$$\widetilde{J}_{\alpha\alpha} = -\,|\mathcal{T}_\alpha|^{-1} \sum_\gamma \left(\mathbf{c} \cdot \boldsymbol{a}_{\alpha\gamma}\right)_+ < 0 \tag{3.18}$$

and the center of each Geršgorin disk of $\widetilde{J}$ lies on the negative real axis. The off diagonal elements of (3.6) satisfy

$$\widetilde{J}_{\alpha\beta} = -\,|\mathcal{T}_\alpha|^{-1} \left(\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta}\right)_- \geq 0\,. \tag{3.19}$$

The radius ρ_α of the α-th Geršgorin disk is thus given by

$$\rho_\alpha = \sum_{\beta \neq \alpha} \left|\widetilde{J}_{\alpha\beta}\right| = \sum_{\beta \neq \alpha} \widetilde{J}_{\alpha\beta} = -\,|\mathcal{T}_\alpha|^{-1} \sum_\beta \left(\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta}\right)_-\ .$$

The geometric relation (2.3) proves that $\rho_\alpha = -\widetilde{J}_{\alpha\alpha}$. Thus all Geršgorin disks lie in the left closed half plane of the complex plane. According to point (i) of Theorem 3.3 the same is true for the spectrum of $\widetilde{J}$. Therefore, the dynamical system (3.4) cannot have exponentially growing solutions. The following theorem gives a complete result for the stability of (3.4) that excludes even polynomially growing solutions and is valid on arbitrary meshes and in arbitrary dimension. An equivalent stability result for the fully discrete finite volume scheme can be found in [17].

Theorem 3.6 (Stability of the First Order Finite-Volume Operator) *Let $\widetilde{J}$ be the first order finite volume operator (3.6) and let $\|.\|$ denote the norm induced by the usual Hermitian product $(.,.)$ on $\mathbb{C}^N$. Then any solution of system (3.4) satisfies*

$$\sup_{t\geq 0} \|\mathfrak{u}\,(t)\| \leq \mathrm{C}\,\|\mathfrak{u}_0\|$$

with C given by

$$\mathrm{C} = \sqrt{\frac{\max_\alpha |\mathcal{T}_\alpha|}{\min_\alpha |\mathcal{T}_\alpha|}}\ .$$

This result holds on arbitrary grids and for all advection velocities $\mathbf{c} \in \mathbb{R}^d$.

Proof : It is sufficient to prove that the operator (3.6) has property (ii) of Theorem 3.4. Some simple identities make this possible. The first is the antisymmetry of the face normal, i.e.

$$(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = (-\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha})_- = -\left(\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha}\right)_+\ . \tag{3.20}$$

The second is the identity (2.3) in the form

$$\sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ + \sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = 0 \tag{3.21}$$

In addition, equations (3.20) and (3.21) provide the useful identity

$$\begin{aligned}\textstyle\sum_\alpha |u|^2_\alpha \sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ &= \textstyle\sum_\alpha |u|^2_\alpha\,(-1)\sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = \\ \textstyle\sum_\alpha |u|^2_\alpha \sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ &= \quad \textstyle\sum_\beta |u|^2_\beta \sum_\alpha (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\end{aligned} \tag{3.22}$$

that can also be found in [17]. Let $\mathfrak{u}$ be a solution of the semi-discrete equation

$$\frac{d\mathfrak{u}\,(t)}{dt} = \widetilde{J}\mathfrak{u}\,(t)\ ,\ \mathfrak{u}\,(0) = \mathfrak{u}_0$$

Consider the positive definite diagonal matrix $\hat{G}$ with elements

$$\hat{g}_{\alpha\beta} = |\mathcal{T}_\alpha|\,\delta_{\alpha\beta} \tag{3.23}$$

Identities (3.20) and (3.22) allow to write $\frac{d}{dt}\left(\mathfrak{u}, \hat{G}\mathfrak{u}\right)$ as

$$\begin{aligned}
&\frac{d}{dt}\left(\mathfrak{u}, \hat{G}\mathfrak{u}\right) = \left(\mathfrak{u}, \left[\hat{G}\tilde{J} + \tilde{J}^*\hat{G}\right]\mathfrak{u}\right) = \qquad (3.24)\\
&= -2\Re\left\{\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ |u|^2_\alpha + \sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- u^*_\alpha u_\beta\right\} =\\
&= -2\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ |u_\alpha|^2 - \sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- \left(u^*_\alpha u_\beta + u_\alpha u^*_\beta\right) =\\
&= -\sum_{\alpha,\beta}\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \left(|u|^2_\alpha + |u|^2_\beta\right) - \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \left(u^*_\alpha u_\beta + u_\alpha u^*_\beta\right)\right\} =\\
&= -\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ |u_\alpha - u_\beta|^2 = -\frac{1}{2}\sum_{\alpha,\beta}\left|\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right| |u_\alpha - u_\beta|^2 \le 0.
\end{aligned}$$

Therefore $\left(\mathfrak{u}(t), \hat{G}\mathfrak{u}(t)\right) \le \left(\mathfrak{u}(0), \hat{G}\mathfrak{u}(0)\right)$ for all $t \ge 0$. The definition of $\hat{G}$ in (3.23) yields the bounds

$$\frac{\left(\mathfrak{u}, \hat{G}\mathfrak{u}\right)}{\max_\alpha |\mathcal{T}_\alpha|} \le (\mathfrak{u}, \mathfrak{u}) \le \frac{\left(\mathfrak{u}, \hat{G}\mathfrak{u}\right)}{\min_\alpha |\mathcal{T}_\alpha|} \text{ for all } \mathfrak{u} \in \mathbb{C}^N.$$

This implies that

$$\sqrt{(\mathfrak{u}(t), \mathfrak{u}(t))} \le \sqrt{\frac{\max_\alpha |\mathcal{T}_\alpha|}{\min_\alpha |\mathcal{T}_\alpha|}}\sqrt{(\mathfrak{u}(0), \mathfrak{u}(0))} \text{ for all } t \ge 0.$$

□

3.4 Stability Analysis of the MUSCL *Scheme : General Setting*

In contrast to the first order operator (3.6), the MUSCL operator (3.7) depends on the slope reconstruction and its stencil. The question of eigenvalue stability of the MUSCL operator (3.7) on irregular grids can be formulated in the following way. Are there *consistent* reconstruction coefficients $\boldsymbol{s}_{\alpha\beta}$ in the sense of (2.19) so that for any velocity $\boldsymbol{c} \in \mathbb{R}^d$ there is a positive definite matrix G with the property that $GJ + J^*G$ is negative semi-definite? The reconstruction should provide linear stability regardless of the convection velocity. This is necessary for applications to gas dynamics where the velocity is not fixed.

Unfortunately, in the case of the MUSCL operator (3.7), the simple diagonal matrix $\hat{G}$ (3.23) with elements $\hat{g}_{\alpha\beta} = |\mathcal{T}_\alpha|\,\delta_{\alpha\beta}$ does no longer give a general stability result. The numerical evidence shows that $\hat{G}J + J^*\hat{G}$ is in general not negative semidefinite on unstructured grids, even when the MUSCL operator (3.7) is stable. It turns out to be very difficult to find a matrix G that could give a rigorous proof of eigenvalue stability for the MUSCL operator (3.7).

Despite this fact, it is interesting and very important to have a strategy at hand that helps to design a MUSCL scheme with the best possible stability properties. Consider the increase of the quadratic energy $\left(\mathrm{u}, \hat{G}\mathrm{u}\right)$ given by

$$\frac{d}{dt}\left(\mathrm{u}, \hat{G}\mathrm{u}\right) = \left(\mathrm{u}, \left[\hat{G}J + J^*\hat{G}\right]\mathrm{u}\right) \tag{3.25}$$

with the diagonal matrix $\hat{G}$ from (3.23). The goal is to look for reconstruction coefficients that make (3.25) as small as possible. This approach does not give a general criterion for the eigenvalue stability of (3.7) but it allows to identify the consistent reconstruction methods that lead to the smallest increase of $\left(\mathrm{u}, \hat{G}\mathrm{u}\right)$.

With the use of the identities (3.20), (3.21) and (3.22), expression (3.25) becomes

$$\begin{aligned}\frac{d}{dt}\left(\mathrm{u}, \hat{G}\mathrm{u}\right) = \sum_{\alpha,\beta} (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \Bigg[&-|u_\beta - u_\alpha|^2 + \\ &+2\sum_{\gamma} \Re\left\{\left(u_\beta^* - u_\alpha^*\right)\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{s}_{\alpha\gamma}\left(u_\gamma - u_\alpha\right)\right\}\Bigg] \ .\end{aligned} \tag{3.26}$$

The form of the quadratic expression (3.26) justifies the introduction of

Definition 3.7 (Local Reconstruction Map) *Let m_α be the number of cells in the reconstruction stencil of cell $\mathcal{T}_\alpha$ and let l_α be its number of first neighbors, i.e. its number of faces. The* local reconstruction map *of cell $\mathcal{T}_\alpha$ is the $l_\alpha \times m_\alpha$ matrix R_α with coefficients $r^{(\alpha)}_{\beta\gamma} = \boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{s}_{\alpha\gamma}$.*

Remark 3.8 The local reconstruction map determines how the fluctuations $u_\gamma - u_\alpha$ translate into the fluctuations $u_{\alpha\beta} - u_\alpha$ where $u_{\alpha\beta}$ is the reconstructed value at the interface between cells $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$. In matrix notation, R_α can be written as $R_\alpha = K_\alpha S_\alpha$ where K_α is the $l_\alpha \times d$ matrix with rows $\boldsymbol{k}_{\alpha\beta}$

$$K^t_\alpha = \left[\boldsymbol{k}_{\alpha\beta_1}, \boldsymbol{k}_{\alpha\beta_2}, \dots, \boldsymbol{k}_{\alpha\beta_{l_\alpha}}\right] \tag{3.27}$$

and $\mathbb{V}_\alpha \triangleq \{\beta_1, \dots, \beta_{l_\alpha}\}$ are the first neighbors of cell $\mathcal{T}_\alpha$. It is important not to confound the local reconstruction matrix $R_\alpha = K_\alpha S_\alpha$ with the slope reconstruction matrix S_α. The matrix R_α has the valuable properties to be dimensionless and invariant under scalings of the mesh. It describes the local reconstruction geometry in cell $\mathcal{T}_\alpha$.

With Definition 3.7 and the notation $\delta u_{\alpha\beta} = u_\beta - u_\alpha$, (3.26) can be written as a sum

$$\frac{d}{dt}\left(\mathrm{u}, \hat{G}\mathrm{u}\right) = \sum_{\alpha=1}^{N} \left(\Theta_\alpha\left(\delta\mathrm{u}\right) + \Phi_\alpha\left(\delta\mathrm{u}\right)\right) \tag{3.28}$$

where $\Theta_\alpha\left(\delta\mathrm{u}\right)$ and $\Phi_\alpha\left(\delta\mathrm{u}\right)$ are defined as

$$\Theta_\alpha\left(\delta\mathrm{u}\right) \triangleq -\sum_\beta \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \left|\delta u_{\alpha\beta}\right|^2 \leq 0 \tag{3.29}$$

$$\Phi_\alpha\left(\delta\mathrm{u}\right) \triangleq 2\sum_\beta\sum_\gamma \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \Re\left\{\delta u^*_{\alpha\beta} r^{(\alpha)}_{\beta\gamma} \delta u_{\alpha\gamma}\right\} \tag{3.30}$$

for each cell $\mathcal{T}_\alpha$.

The Definitions (3.29) and (3.30) show clearly that any increase of the expression $\left(\mathrm{u}, \hat{G}\mathrm{u}\right)$ can only come from the terms $\Phi_\alpha\left(\delta\mathrm{u}\right)$ in (3.28). Furthermore, $\Phi_\alpha\left(\delta\mathrm{u}\right)$ depends linearly on the entries $r^{(\alpha)}_{\beta\gamma}$ of the local reconstruction map R_α in each cell. It is also very important to note that the reconstruction matrix R_α of cell $\mathcal{T}_\alpha$ and with it the reconstruction coefficients $\boldsymbol{s}_{\alpha\gamma}$ in cell $\mathcal{T}_\alpha$ occur exclusively in the term Φ_α associated with cell $\mathcal{T}_\alpha$. In this way the reconstruction coefficients that minimize (3.28) can be chosen by minimizing the Φ_α separately for each cell. Before the presentation of the main results in Subsection 3.6, we give a brief overview of the one-dimensional case.

3.5 Stability Analysis of the MUSCL *Scheme : the One-Dimensional Case.*

This subsection contains a short treatment of the one-dimensional case. Its subject is the stability analysis of the MUSCL scheme in one space dimension when it is applied to the convection equation

$$\partial_t u + c\partial_x u = 0\,;\; c > 0 \tag{3.31}$$

on an irregular periodic grid of size N. For this subsection only, we adopt a specific notation that better suits the one-dimensional case, see Fig. 3.1. The cells are indexed by $j \in \{1, \ldots, N\}$ with the convention that $N + 1 = 1$ to handle the periodicity of the grid. In cell $\mathcal{T}_j$ we define a dimensionless volume $\alpha_j > 0$ by the relation $\alpha_j = \frac{1}{h}\left|\mathcal{T}_j\right|$ where h is a length scale. In one space dimension, the distance between the barycenters of the adjacent

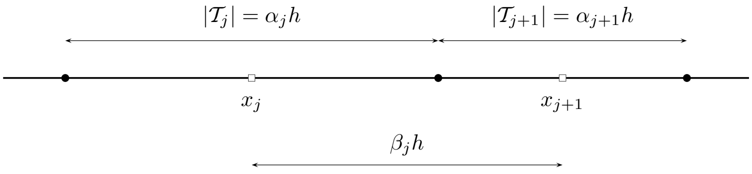

$|\mathcal{T}_j| = \alpha_j h$
$|\mathcal{T}_{j+1}| = \alpha_{j+1} h$
x_j
x_{j+1}
$\beta_j h$

Figure 3.1: The cells $\mathcal{T}_j$, $\mathcal{T}_{j+1}$, with barycenters x_j, x_{j+1} and lengths $\alpha_j h$, $\alpha_{j+1} h$.

cells $\mathcal{T}_j$ and $\mathcal{T}_{j+1}$ is $h_{j,j+1} = \frac{1}{2}\left(|\mathcal{T}_j| + |\mathcal{T}_{j+1}|\right)$. A dimensionless distance β_j is defined by $\beta_j = \frac{1}{h} h_{j,j+1}$. The β_j are related to the α_j by the formula

$$\beta_j = \frac{1}{2}\left(\alpha_j + \alpha_{j+1}\right)$$

as depicted in Fig. 3.1. In particular the relations $h\beta_{j-1} = h_{j-1,j} = -h_{j,j-1}$ and $\beta_j > 0$ hold for all j.

The slope reconstruction matrix defined by (2.21) for reconstruction on the first neighborhood is the 1×2 matrix

$$S_j = \begin{bmatrix} s_{j,j+1} \; s_{j,j-1} \end{bmatrix} . \tag{3.32}$$

In one space dimension, the distances between the barycenter of cell $\mathcal{T}_j$ and the barycenters of its left and right faces can be written as functions of the cell volume

$$k_{j,j+1} = \frac{|\mathcal{T}_j|}{2} = \frac{h\alpha_j}{2} \, , \; k_{j,j-1} = -\frac{|\mathcal{T}_j|}{2} = -\frac{h\alpha_j}{2} \, . \tag{3.33}$$

In this context, the matrix R_j of the local reconstruction map from Definition 3.7 becomes

$$R_j = \begin{bmatrix} k_{j,j+1} s_{j,j+1} & k_{j,j+1} s_{j,j-1} \\ k_{j,j-1} s_{j,j+1} & k_{j,j-1} s_{j,j-1} \end{bmatrix} . \tag{3.34}$$

In the one-dimensional case, the introduction of the coefficients r_j^+ and r_j^- simplifies the notation as follows

$$\begin{array}{ll} r_j^+ = k_{j,j+1} s_{j,j+1} = & \frac{|\mathcal{T}_j|}{2} s_{j,j+1} \; = \; \frac{h\alpha_j}{2} s_{j,j+1} \\ r_j^- = k_{j,j-1} s_{j,j-1} = & -\frac{|\mathcal{T}_j|}{2} s_{j,j-1} = -\frac{h\alpha_j}{2} s_{j,j-1} \end{array} . \tag{3.35}$$

This allows to express the reconstructed values at the cell interfaces as

$$\begin{array}{l} u_j^+ = u_{j,j+1} = u_j + r_j^+ (u_{j+1} - u_j) + r_j^- (u_j - u_{j-1}) \\ u_j^- = u_{j,j-1} = u_j - r_j^+ (u_{j+1} - u_j) - r_j^- (u_j - u_{j-1}) \end{array} . \tag{3.36}$$

With the help of this notation, the one-dimensional semi-discrete MUSCL scheme in (3.5) is reduced to the equation

$$\frac{du_j\left(t\right)}{dt} = -\frac{c}{|\mathcal{T}_j|}\left(u_j^+\left(t\right) - u_{j-1}^+\left(t\right)\right) . \tag{3.37}$$

The consistency relation (2.19) for the slope reconstruction becomes

$$s_{j,j+1}h_{j,j+1} + s_{j,j-1}h_{j,j-1} = 1 \tag{3.38}$$

In the one-dimensional notation, equation (3.38) translates into

$$\beta_j r_j^+ + \beta_{j-1} r_j^- = \frac{1}{2}\alpha_j . \tag{3.39}$$

Two commonly used slope reconstructions are the least squares reconstruction

$$r_j^+ = \frac{1}{2}\frac{\alpha_j\beta_j}{\beta_j^2 + \beta_{j-1}^2} , \; r_j^- = \frac{1}{2}\frac{\alpha_j\beta_{j-1}}{\beta_j^2 + \beta_{j-1}^2} \tag{3.40}$$

and the Green reconstruction

$$r_j^+ = \frac{1}{4}\frac{\alpha_j}{\beta_j} , \; r_j^- = \frac{1}{4}\frac{\alpha_j}{\beta_{j-1}} . \tag{3.41}$$

See Proposition 2.4 for the least squares reconstruction and Subsection 4.3 for the Green reconstruction.

The matrix of the semi-discrete MUSCL operator J defined by (3.7) has four non zero entries in each line. They are given by

$$\begin{array}{rcc} J_{j,j+1} = & & -\frac{c}{\alpha_j h} r_j^+ \\ J_{j,j} = & & -\frac{c}{\alpha_j h}\left(1 - r_j^+ + r_j^- - r_{j-1}^+\right) \\ J_{j,j-1} = & & -\frac{c}{\alpha_j h}\left(-r_j^- - 1 + r_{j-1}^+ - r_{j-1}^-\right) \\ J_{j,j-2} = & & -\frac{c}{\alpha_j h} r_{j-1}^- \end{array} . \tag{3.42}$$

In the one-dimensional case, the surface vectors $\boldsymbol{a}_{j,j+1}$ and $\boldsymbol{a}_{j,j-1}$ are real numbers with the respective values 1 and -1 and the velocity vector $\boldsymbol{c}$ is reduced to a real number c. The assumption $c > 0$ implies $\left(\mathbf{c}\cdot\boldsymbol{a}_{j,j+1}\right)_+ = c$ and $\left(\mathbf{c}\cdot\boldsymbol{a}_{j,j-1}\right)_+ = 0$ and the quadratic form (3.28) becomes

$$\begin{aligned} \frac{d}{dt}\left(\mathrm{u}, \hat{G}\mathrm{u}\right) = c \Bigg(& -\sum_{j=1}^{N}\left(u_{j+1} - u_j\right)^2 + \\ & +2\sum_{j=1}^{N}\left(u_{j+1} - u_j\right)\left(r_j^+\left(u_{j+1} - u_j\right) + r_j^-\left(u_j - u_{j-1}\right)\right)\Bigg) . \end{aligned} \tag{3.43}$$

The functional (3.43) can also be expressed by

$$\frac{d}{dt}\left(\mathrm{u},\hat{G}\mathrm{u}\right)=(Q\mathrm{u},\mathrm{u}) \tag{3.44}$$

where Q is derived from J by

$$Q=\hat{G}J+J^{*}\hat{G} \tag{3.45}$$

and $\hat{G}=\mathrm{diag}(\alpha_1,...,\alpha_N)$ is the diagonal matrix of the dimensionless volumes. A sufficient but not necessary condition for stability is given by

$$\frac{d}{dt}\left(\mathrm{u},\hat{G}\mathrm{u}\right)=(Q\mathrm{u},\mathrm{u})\leq 0\,. \tag{3.46}$$

Due to the symmetry of the matrix Q, the condition (3.46) is equivalent to the condition $\sigma\left(Q\right)\subset\mathbb{C}^{-}$ for the spectrum $\sigma\left(Q\right)$ of Q.

The case of an equidistant grid is easy to handle, since the spectrum is available explicitly by a Discrete Fourier Transform.

Proposition 3.9 (Stability of the MUSCL Scheme on Equidistant Grids) *Consider the linear system (3.37) with velocity $c>0$ on an equidistant grid with the same slope reconstruction method in each cell. In this case, the reconstruction coefficients are the same for all grid cells : $r_j^{+}=r^{+}$ and $r_j^{-}=r^{-}$ for $1\leq j\leq N$. If the two reconstruction coefficients satisfy the consistency condition*

$$r^{+}+r^{-}=\frac{1}{2} \tag{3.47}$$

then the real and imaginary parts of the eigenvalues of the MUSCL *operator J (3.42) are given explicitly by*

$$\begin{array}{rr}\Re\left(\lambda_k\right)= & -\frac{2c}{h}r^{-}\left(1-\cos\left(\frac{2\pi k}{N}\right)\right)^2\\ \Im\left(\lambda_k\right)= & -\frac{c}{h}\sin\left(\frac{2\pi k}{N}\right)\left[1+2r^{-}\left(1-\cos\left(\frac{2\pi k}{N}\right)\right)\right]\end{array} \tag{3.48}$$

where $-\frac{N}{2}<k\leq\left[\frac{N}{2}\right]$. This shows that in the equidistant setting with a positive velocity $c>0$ the one-dimensional MUSCL *scheme is stable if and only if $r^{-}\geq 0$.*

□

Remark 3.10 Note that for equidistant meshes the stability of the semi-discrete MUSCL scheme (3.37) can also be deduced from the sufficient condition (3.46) by means of the Geršgorin location of the eigenvalues, see Theorem 3.3. Furthermore, the stability result of Proposition 3.9 can be extended to multidimensional cartesian grids by the use of Kronecker products, [28, Chapter 4].

In the case of an irregular periodic grid the full stability analysis is a difficult task, since it amounts to prove that the spectrum of J lies in the left-half complex plane. We limit ourselves to the following numerical observations.

- The numerical computation of the spectrum of the matrix J defined by (3.42) has been performed with the least square reconstruction (3.40) and with the Green reconstruction (3.41) for a series of grids with random cell length from $N = 4$ to $N = 250$. The spectral abscissa $\max(\Re(\lambda))$ has been shown to satisfy the stability condition $\max \Re(\lambda) \leq 0$ up to the numerical accuracy of the computer. This allows to infer that the semi-discrete MUSCL scheme with these two slope reconstructions is stable regardless of the shape of the grid. The computation of the spectra has been performed independently in MATLAB and in MAPLE.
- On the contrary, the sufficient stability condition (3.46) is clearly violated on numerous simple examples. This means that the energy

  $$E(t) = \sum_{j=1}^{N} |\mathcal{T}_j| \, |u_j(t)|^2$$

  can grow locally for some solution $t \mapsto u(t)$ even if the scheme (3.37) is stable. In other words, the standard energy $E(t)$ is actually not sufficient to assess the stability of the scheme. This is a classical drawback of any energy method : in the absence of knowledge of the correct energy function one cannot conclude.

3.6 Minimization Properties of the Least-Squares Method

In this section, we return to unstructured grids in arbitrary dimension and ask if it is possible to identify a criterion for the reconstruction coefficients such that the increase of the energy given by (3.28)

$$\frac{d}{dt}\left(\mathrm{u}, \hat{G}\mathrm{u}\right) = \sum_{\alpha=1}^{N} \left(\Theta_\alpha(\delta \mathrm{u}) + \Phi_\alpha(\delta \mathrm{u})\right)$$

becomes small. The discussion at the end of section 3.4 shows that any increase of $\left(\mathrm{u}, \hat{G}\mathrm{u}\right)$ can only be caused by the terms $\Phi_\alpha(\delta \mathrm{u})$ in (3.28). Furthermore, the terms $\Theta_\alpha(\delta \mathrm{u})$ in (3.28) are independent of the slope reconstruction and the terms (3.30)

$$\Phi_\alpha(\delta \mathrm{u}) \triangleq 2 \sum_{\beta} \sum_{\gamma} (\mathbf{c} \cdot \boldsymbol{a}_{\alpha\beta})_+ \Re \left\{ \delta u^*_{\alpha\beta} r^{(\alpha)}_{\beta\gamma} \delta u_{\alpha\gamma} \right\}$$

depend linearly on the entries $r^{(\alpha)}_{\beta\gamma}$ of the local reconstruction map R_α of the cell $\mathcal{T}_\alpha$.

The general form (2.23) of the consistent slope reconstruction shows that the local reconstruction map can be written as

$$R_\alpha = K_\alpha S_\alpha = K_\alpha \left(\tilde{S}_\alpha + \varLambda_\alpha B_\alpha \right) \tag{3.49}$$

where $\tilde{S}_\alpha = \left(H_\alpha^t H_\alpha \right)^{-1} H_\alpha^t$ is the least squares reconstruction matrix and $\varLambda_\alpha$ is an arbitrary parameter matrix that represents the degrees of freedom of the consistent slope reconstruction. Recall that the matrix B_α is a maximum rank solution of (2.24) that satisfies in particular $B_\alpha \tilde{S}_\alpha^t = 0$ and $\tilde{S}_\alpha B_\alpha^t = 0$.

Equation (3.49) suggests to choose in each cell $\mathcal{T}_\alpha$ the matrix $\varLambda_\alpha$ in such a way that the term $\varPhi_\alpha \left(\delta \mathrm{u} \right)$ becomes as small as possible. However, a direct minimization of (3.30) for fixed values of $\delta \mathrm{u}$ and $\boldsymbol{c}$ encounters two problems.

1. First, it is not obvious that the infimum of $\varPhi_\alpha \left(\delta \mathrm{u} \right)$ with regard to $\varLambda_\alpha$ exists because $\varPhi_\alpha \left(\delta \mathrm{u} \right)$ may not be bounded from below as a function of $\varLambda_\alpha$.
2. Second, if there is a matrix $\varLambda_\alpha$ that minimizes $\varPhi_\alpha \left(\delta \mathrm{u} \right)$ for given values $\delta \mathrm{u}$ and $\boldsymbol{c}$ then $\varLambda_\alpha$ depends itself on these specific values of the solution $\delta \mathrm{u}$ and the velocity $\boldsymbol{c}$.

To overcome these obstacles, we propose an alternative criterion for the slope reconstruction that makes an *approximate and qualitative statement* about the influence of the slope reconstruction on the increase of the energy $\left(\mathrm{u}, \hat{G} \mathrm{u} \right)$. This criterion is based on matrix norms of the local reconstruction map. We stress that we use the term *matrix norm* for any norm on the space of complex matrices $\mathbb{M}_{k,n} \left(\mathbb{C} \right)$ seen as a linear vector space. This means that the norms used hereafter do not necessarily satisfy the estimate $\| AB \| \leq \| A \| \, \| B \|$.

The criterion is given by the

Definition 3.11 (Optimal Consistent Slope Reconstructions) *Let* $\| . \|_{\mathrm{M}}$ *be a matrix norm on the space of matrices* $\mathbb{M}_{k,n} \left(\mathbb{C} \right)$. *A consistent slope reconstruction given by a matrix* $\breve{S}_\alpha$ *is called* optimal with regard to the norm $\| . \|_{\mathrm{M}}$ *if it is a solution of the minimization problem*

$$\left\| K_\alpha \breve{S}_\alpha \right\|_{\mathrm{M}} = \min \left\{ \left\| K_\alpha S_\alpha \right\|_{\mathrm{M}} \middle| \, S_\alpha H_\alpha = \boldsymbol{I}_{d \times d} \, , \, S_\alpha \in \mathbb{M}_{d,m} \left(\mathbb{R} \right) \right\} . \tag{3.50}$$

For any matrix norm, the minimization problem (3.50) of Definition 3.11 is a problem of convex optimization. Indeed, the function

$$\mathbb{M}_{d,m-d} \left(\mathbb{C} \right) \ni \varLambda_\alpha \longmapsto \left\| K_\alpha \left(\tilde{S}_\alpha + \varLambda_\alpha B_\alpha \right) \right\|_{\mathrm{M}}$$

is a continuous convex function that is bounded from below. Furthermore, any minimizer Λ_α of $\|K_\alpha S_\alpha\|_{\mathrm{M}}$ is independent of the solution $\delta\mathfrak{u}$ and the velocity $\boldsymbol{c}$. Therefore, this approach seems better suited for a generalization to applications where the fluid velocity is not fixed. It is also noteworthy that this minimization problem is invariant under scalings of the grid. The importance of this criterion is underlined by the

Remark 3.12 Consider the criterion of Definition 3.11 for the spectral norm $\|.\|_2$. Denote by $\mathbb{V}_\alpha$ the set of cell indexes of the direct neighbors and by $\mathbb{W}_\alpha$ the set of cell indexes in the reconstruction stencil. Furthermore, denote by $u_{\alpha\beta}$ the reconstructed values at the cell interfaces. If $\|K_\alpha S_\alpha\|_2 > 1$, then there exists a $\mathfrak{u} = (u_1, \ldots, u_N)$ such that

$$\sum_{\beta\in\mathbb{V}_\alpha} |u_{\alpha\beta} - u_\alpha|^2 > \sum_{\beta\in\mathbb{W}_\alpha} |u_\beta - u_\alpha|^2 \;.$$

This means that the fluctuations between the values at the faces and at the center of the cell become larger in the quadratic mean than the fluctuations between the cell neighbors. Such an undesirable behaviour has been observed for the first neighborhood reconstruction on tetrahedral grids, see Section 4.5. This finding is a motivation to make $\|K_\alpha S_\alpha\|_2$ as small as possible. □

However, the new criterion has the problem that any minimizer Λ_α of $\|K_\alpha S_\alpha\|_{\mathrm{M}}$ depends on the choice of the matrix norm $\|.\|_{\mathrm{M}}$. For this reason, the new criterion seems only pertinent if there is a common minimizer Λ_α of $\|K_\alpha S_\alpha\|_{\mathrm{M}}$ for at least a family of norms. An important family of matrix norms is given by the

Definition 3.13 (Unitarily Invariant Matrix Norms) *A matrix norm* $\|.\|_{\mathrm{M}}$ *is called* unitarily invariant *if* $\|UAV\|_{\mathrm{M}} = \|A\|_{\mathrm{M}}$ *for all matrices* $A \in \mathbb{M}_{k,n}(\mathbb{C})$ *and all unitary matrices* $U \in \mathbb{M}_k(\mathbb{C})$ *and* $V \in \mathbb{M}_n(\mathbb{C})$.

The family of unitarily invariant matrix norms includes the Frobenius norm, the spectral norm, the trace norm, the Ky Fan norms etc., see [28]. An example of a matrix norm that is not unitarily invariant is given by

$$\begin{aligned} \|A\|_{\mathcal{L}(2,\infty)} &\triangleq \sup_{1\le\alpha\le l}\ \sup_{\|\mathfrak{z}\|=1} \left| \sum_{1\le\beta\le m} a_{\alpha\beta} z_\beta \right| \\ &= \sup_{1\le\alpha\le l} \sqrt{\sum_{1\le\beta\le m} |a_{\alpha\beta}|^2}\;. \end{aligned} \tag{3.51}$$

These definitions allow to prove the

Theorem 3.14 (Minimization Property of the Least Squares Reconstruction) *Consider the local reconstruction map $R_\alpha = K_\alpha S_\alpha$ in cell $\mathcal{T}_\alpha$ where the slope reconstruction matrix S_α satisfies the consistency condition (2.22). Then the least squares reconstruction matrix $\tilde{S}_\alpha$ minimizes the following functions of $R_\alpha = K_\alpha S_\alpha$ under the constraint of consistency $S_\alpha H_\alpha = \boldsymbol{I}_{d\times d}$.*

(i) $\tilde{S}_\alpha$ minimizes each of the singular values of $K_\alpha S_\alpha$.

(ii) $\tilde{S}_\alpha$ minimizes all unitarily invariant matrix norms of $K_\alpha S_\alpha$. In particular, $\tilde{S}_\alpha$ minimizes the spectral norm, the Frobenius norm, the trace norm and all Ky Fan norms of $K_\alpha S_\alpha$.

(iii) $\tilde{S}_\alpha$ minimizes the norm (3.51) of $K_\alpha S_\alpha$. It minimizes further any matrix norm of $K_\alpha S_\alpha$ that can be expressed as $\|A\|_M = F(AA^*)$ where F is a function of Hermitian matrices such that $F(P) \leq F(P+Q)$ for all Hermitian P and all positive semidefinite Q.

Proof : As explained in Section 2.2, a general consistent reconstruction can be written as $S_\alpha = \tilde{S}_\alpha + \Lambda_\alpha B_\alpha$ where $\tilde{S}_\alpha$ is the least squares reconstruction matrix, the matrix B_α is a maximal rank solution of $B_\alpha H_\alpha = 0$ and the matrix Λ_α represents the degrees of freedom of the consistent reconstruction. The singular values of $K_\alpha S_\alpha$ are the square roots of the eigenvalues of the matrix

$$K_\alpha S_\alpha S_\alpha^t K_\alpha^t = \left(K_\alpha \tilde{S}_\alpha + K_\alpha \Lambda_\alpha B_\alpha\right)\left(\tilde{S}_\alpha^t K_\alpha^t + B_\alpha^t \Lambda_\alpha^t K_\alpha^t\right) \tag{3.52}$$

The least squares reconstruction matrix is given by $\tilde{S}_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$ and fulfills therefore $\tilde{S}_\alpha B_\alpha^t = 0$. Consequently, the matrix (3.52) is a sum of two positive semidefinite matrices

$$K_\alpha S_\alpha S_\alpha^t K_\alpha^t = K_\alpha \tilde{S}_\alpha \tilde{S}_\alpha^t K_\alpha^t + K_\alpha \Lambda_\alpha B_\alpha B_\alpha^t \Lambda_\alpha^t K_\alpha^t\,. \tag{3.53}$$

The proof of item (i) uses Corollary 4.4.3 on page 182 in [27] : Let P and Q be Hermitian matrices and let Q be positive semidefinite. Assume that the eigenvalues of P and $P+Q$ are arranged in increasing order and denote by $\lambda_k(P)$ and $\lambda_k(P+Q)$ the k-th eigenvalues of P and $P+Q$. Then $\lambda_k(P) \leq \lambda_k(P+Q)$. This proves that the k-th eigenvalues of the matrices in (3.53) satisfy

$$\lambda_k\left(K_\alpha \tilde{S}_\alpha \tilde{S}_\alpha^t K_\alpha^t\right) \leq \lambda_k\left(K_\alpha \tilde{S}_\alpha \tilde{S}_\alpha^t K_\alpha^t + K_\alpha \Lambda_\alpha B_\alpha B_\alpha^t \Lambda_\alpha^t K_\alpha^t\right)\,. \tag{3.54}$$

The vector of the singular values of $K_\alpha S_\alpha$ is the vector of the square roots of the eigenvalues of $K_\alpha S_\alpha S_\alpha^t K_\alpha^t$ arranged in increasing order. Therefore (3.54) shows that the k-th singular value of $K_\alpha S_\alpha$ has a minimum at $\Lambda_\alpha = 0$, i.e. at the least squares reconstruction matrix. This proves the first item of Theorem 3.14.

The proof of the second item (ii) makes use of Definition 3.5.17 on page 209 and Theorem 3.5.18 on page 210 in [28]. For any unitarily invariant matrix norm, and in particular for the norms cited above, there is a symmetric gauge function such that the norm can be expressed as the gauge function of the vector of singular values. A symmetric gauge function is moreover a monotone vector norm. As the least squares reconstruction minimizes each entry of the vector of singular values of $K_\alpha S_\alpha$ among the consistent reconstructions, it minimizes also the gauge function of the singular values and by consequence the matrix norm.

The proof of the third item proceeds in the same way. The norm (3.51) can be written as

$$\|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)} = \sup_{\beta\in\mathbb{V}_\alpha} \sup_{\|\mathfrak{z}\|=1} \left| \sum_\gamma \boldsymbol{k}_{\alpha\beta} \cdot \boldsymbol{s}_{\alpha\gamma} z_\gamma \right| \tag{3.55}$$

$$= \sup_{\beta\in\mathbb{V}_\alpha} \sqrt{\sum_{1\leq\gamma\leq m_\alpha} \left(K_\alpha \tilde{S}_\alpha + K_\alpha \Lambda_\alpha B_\alpha\right)_{\beta\gamma} \left(\tilde{S}^t_\alpha K^t_\alpha + B^t_\alpha \Lambda^t_\alpha K^t_\alpha\right)_{\gamma\beta}} \, .$$

As shown in the proof of the first item, expression (3.55) takes the following form

$$\|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)} = \sup_{\beta\in\mathbb{V}_\alpha} \sqrt{\left(K_\alpha \tilde{S}_\alpha \tilde{S}^t_\alpha K^t_\alpha\right)_{\beta\beta} + (K_\alpha \Lambda_\alpha B_\alpha B^t_\alpha \Lambda^t_\alpha K^t_\alpha)_{\beta\beta}}$$

that has a minimum at $\Lambda_\alpha = 0$, that is at the least squares reconstruction matrix. This proves the minimization property for the norm (3.51). Now consider any matrix norm of the form $\|A\|_M = F\,(AA^*)$ with a function F having the property stated in item (iii). The same argument as before shows that

$$\left\|K_\alpha \tilde{S}_\alpha\right\|_M = F\left(K_\alpha \tilde{S}_\alpha \tilde{S}^t_\alpha K^t_\alpha\right) \leq F\left(K_\alpha \tilde{S}_\alpha \tilde{S}^t_\alpha K^t_\alpha + K_\alpha \Lambda_\alpha B_\alpha B^t_\alpha \Lambda^t_\alpha K^t_\alpha\right)$$

which proves that the matrix norm in question has a minimum at the least squares reconstruction.

□

The Theorem 3.14 shows that the least squares reconstruction is a common minimizer of the norm of the local reconstruction map for a certain family of matrix norms. For these norms, the least squares reconstruction turns out to be optimal in the sense of the Definition 3.11.

The next step is to investigate the influence of the reconstruction stencil on the local reconstruction map and on the stability of the MUSCL operator (3.7). The following result shows the influence of an extension of the reconstruction stencil on the norms of the local reconstruction map $K_\alpha S_\alpha$ in the case of the least squares reconstruction.

Theorem 3.15 (Influence of the Stencil on the Least Squares Reconstruction) *Consider a fixed reconstruction stencil in cell $\mathcal{T}_\alpha$ and denote by H_α the corresponding geometric matrix (2.20) of rank d. Let $\tilde{S}_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$ denote the slope reconstruction matrix of the least squares method in cell $\mathcal{T}_\alpha$ for that stencil.*

Consider an extension of the reconstruction stencil of cell $\mathcal{T}_\alpha$ by adding a number $l \geq 1$ of cells with indexes $\{\beta_1, \ldots, \beta_l\}$ to the stencil. Let $\check{H}_\alpha$ denote the $l \times d$ matrix whose rows are the l new vectors $\{\boldsymbol{h}_{\alpha\beta_1}, \ldots, \boldsymbol{h}_{\alpha\beta_l}\}$ defined by $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha$ so that the new geometric matrix is given by

$$\hat{H}_\alpha^t = \left[H_\alpha^t \middle| \, \check{H}_\alpha^t\right] .$$

Denote by $\hat{S}_\alpha = \left(\hat{H}_\alpha^t \hat{H}_\alpha\right)^{-1} \hat{H}_\alpha^t$ the matrix of the least squares slope reconstruction on the extended neighborhood. Then the following results hold.

(i) The singular values of $K_\alpha \hat{S}_\alpha$ and $K_\alpha \tilde{S}_\alpha$ satisfy the estimates

$$\sigma_j\left(K_\alpha \hat{S}_\alpha\right) \leq \sigma_j\left(K_\alpha \tilde{S}_\alpha\right) , 1 \leq j \leq d .$$

Let $\|.\|_M$ be any unitarily invariant matrix norm, the norm (3.51) or any norm that can be written as $\|A\|_M = F\left(AA^*\right)$ where F is a function of Hermitian matrices satisfying $F\left(P\right) \leq F\left(P+Q\right)$ for all Hermitian P and all positive semidefinite Q. Then $K_\alpha \hat{S}_\alpha$ and $K_\alpha \tilde{S}_\alpha$ satisfy the estimate

$$\left\|K_\alpha \hat{S}_\alpha\right\|_M \leq \left\|K_\alpha \tilde{S}_\alpha\right\|_M .$$

(ii) Suppose further that the matrix $\check{H}_\alpha$ has full rank d. If $\sigma_j\left(K_\alpha \tilde{S}_\alpha\right) > 0$ for any $1 \leq j \leq d$, then

$$\sigma_j\left(K_\alpha \hat{S}_\alpha\right) < \sigma_j\left(K_\alpha \tilde{S}_\alpha\right) .$$

Let $\|.\|_M$ be any matrix norm cited in item (i). For the norms that can be written in the form $\|A\|_M = F\left(AA^*\right)$ suppose that the strict estimate $F\left(P\right) < F\left(P+Q'\right)$ holds for all Hermitian P and all positive definite Q'. Then $K_\alpha \hat{S}_\alpha$ and $K_\alpha \tilde{S}_\alpha$ satisfy the strict estimate

$$\left\|K_\alpha \hat{S}_\alpha\right\|_M < \left\|K_\alpha \tilde{S}_\alpha\right\|_M .$$

Proof : The proof is based on the Sherman-Morrison-Woodbury matrix identity, see for example [21, page 3] or [30, page 124]. Let A, U, C and V denote complex matrices of the respective sizes $n \times n$, $n \times k$, $k \times k$ and $k \times n$. Then

$$\left(A + UCV\right)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}$$

provided that the inverted matrices exist.

First, we prove properties (i) and (ii) for the singular values. In the case of the least squares reconstruction, the singular values of the matrix $K_\alpha \tilde{S}_\alpha$ are given by the square roots of the eigenvalues of the matrix

$$K_\alpha \tilde{S}_\alpha \tilde{S}_\alpha^t K_\alpha^t = K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t \,. \tag{3.56}$$

It is therefore sufficient to prove each statement for the eigenvalues of the real symmetric matrix (3.56).

If $\check{H}_\alpha$ is the matrix whose rows are the new vectors $\{\boldsymbol{h}_{\alpha\beta_1}, \ldots, \boldsymbol{h}_{\alpha\beta_l}\}$, then the matrix $\left(H_\alpha^t H_\alpha\right)^{-1}$ in (3.56) is replaced by

$$\left(\hat{H}_\alpha^t \hat{H}_\alpha\right)^{-1} = \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} \,. \tag{3.57}$$

The matrix K_α stays the same because the number of faces is unchanged. Application of the Woodbury matrix identity to (3.57), followed by left multiplication by K_α and right multiplication by K_α^t gives the relation

$$\begin{aligned} K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t - K_\alpha \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} K_\alpha^t &= \\ = K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t \left(\boldsymbol{I} + \check{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t\right)^{-1} \check{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t \,. \end{aligned} \tag{3.58}$$

The rest of the proof proceeds in the same way as the proof of Theorem 3.14. The application of Corollary 4.4.3 on page 182 in [27] to equation (3.58) shows that the k-th eigenvalues of the matrices in (3.58) satisfy

$$\lambda_k \left(K_\alpha \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} K_\alpha^t\right) \le \lambda_k \left(K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t\right) \,. \tag{3.59}$$

If $\check{H}_\alpha$ has full rank d, the matrix on the right hand side of (3.58) is positive definite on the orthogonal complement of the null space of K_α^t because the matrix

$$\left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t \left(\boldsymbol{I} + \check{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t\right)^{-1} \check{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1}$$

is in this case positive definite on $\mathbb{R}^d$. All three matrices in (3.58) share the same null space that is equal to the null space of K_α^t. It is now sufficient to apply Weyl's Theorem, see [27, 4.4.1 on page 181], to the restriction of the Hermitian matrices in (3.58) to the orthogonal complement of the null space of K_α^t. This proves that

$$\lambda_k \left(K_\alpha \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} K_\alpha^t\right) < \lambda_k \left(K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t\right) \tag{3.60}$$

for all eigenvalues λ_k that are strictly positive. This argument demonstrates the statement (ii) for the singular values.

The proof for the singular values implies immediately that properties (i) and (ii) hold for all unitarily invariant norms as shown in Theorem 3.14. Now suppose that a matrix norm can be expressed as $\|A\|_M = F(AA^*)$ where F is a function of Hermitian matrices such that $F(P) \leq F(P+Q)$ for Hermitian P and positive semidefinite Q and $F(P) < F(P+Q)$ for Hermitian P and positive definite Q. Then properties (i) and (ii) follow directly from equation (3.58).

Finally, in the case of the least squares reconstruction, the norm (3.51) is given by the formula

$$\left\| K_\alpha \tilde{S}_\alpha \right\|_{\mathcal{L}(\infty,2)} = \sup_{\beta \in \mathbb{V}_\alpha} \sqrt{\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta}}\,. \tag{3.61}$$

According to the Woodbury matrix identity, the equation

$$\begin{aligned} &\left(H_\alpha^t H_\alpha\right)^{-1} - \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} = \\ &= \left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t \left(\boldsymbol{I} + \check{H}_\alpha^t \left(H_\alpha^t H_\alpha\right)^{-1} \check{H}_\alpha^t\right)^{-1} \check{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} \end{aligned} \tag{3.62}$$

holds. This proves statement (i) because

$$\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \leq \boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \tag{3.63}$$

in (3.61). If $\check{H}_\alpha$ has full rank d the matrix on the right hand side of equation (3.62) becomes positive definite. This implies

$$\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha + \check{H}_\alpha^t \check{H}_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} < \boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \tag{3.64}$$

for all vectors $\boldsymbol{k}_{\alpha\beta}$ which proves the statement (ii) for the norm (3.51).

□

3.7 Practical Conclusions

The Definition 3.11 introduces a new criterion to evaluate the impact of the piecewise linear slope reconstruction on the stability of the MUSCL scheme. This criterion defines an *approximate and qualitative measure* to identify the reconstruction methods that are best suited to give a stable MUSCL operator (3.7). The criterion uses a local property in each cell, the local reconstruction map given by Definition 3.7. The subsequent analysis and the theorems of Subsection 3.6 provide two practical conclusions for the choice and design of reconstruction methods and their stencils.

1. Theorem 3.14 shows that the least squares slope reconstruction is a minimizer of the specific criterion of Definition 3.11 in a certain family of norms. This result suggests in particular that if the least squares slope reconstruction gives an unstable scheme, then any other consistent slope reconstruction is also very likely to lead to an unstable scheme.
2. The result of Theorem 3.15 suggests that a larger reconstruction stencil should lead to a more robust scheme. In section 4, this hypothesis is tested numerically and it turns out to be particularly true for three-dimensional meshes.

The purpose of Section 4 is to support and complement these conclusions by explicit numerical computation of spectra.

4 Computation of Spectra of MUSCL Operators in Two and Three Dimensions

The theoretical results of Subsection 3.6 lead to practical recommendations for the slope reconstruction of the MUSCL scheme. The purpose of this section is to underpin and complete these conclusions by the numerical computation of spectra of operator (3.7) on a range of unstructured and structured meshes. One of the main goals is to provide evidence for a relationship between the local reconstruction map given by Definition 3.7 and the eigenvalue stability of the MUSCL operator (3.7).

4.1 Description of the Test Cases

To isolate the influence of the mesh type, the reconstruction method and the reconstruction stencil on the stability of the MUSCL operator (3.7) it is suitable to consider the linear advection equation (3.1) in the simplest setting, i.e. on a square in two dimensions and on a cube in three dimensions with periodic boundary conditions. For each test case, a program written in MAPLE constructs the matrix of the operator J in (3.7) and computes its spectrum by standard eigenvalue algorithms. In addition, it determines for each cell $\mathcal{T}_\alpha$ the value of $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = \|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)}$ and computes the average, median, minimum and maximum values and the 90th, 95th and 99th percentile of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ for each mesh.

The numerical criterion to identify unstable discretizations is the spectral abscissa of J, defined by

$$\omega_J = \max\left\{\Re\left(\lambda\right) \middle|\, \lambda \in \sigma\left(J\right)\right\}.$$

According to proposition (3.1), it is necessary for stability that $\omega_J \leq 0$. Whenever this is false, the operator is unstable in the sense of Definition 3.2.

In two dimensions, the test cases comprise four purely triangular meshes, four hybrid meshes with triangular and quadrilateral elements and four meshes obtained by deformation of uniform cartesian grids. In three dimensions, the test cases consist of four purely tetrahedral meshes, four hybrid grids with tetrahedra, pyramids and prismatic layers at the boundaries and four deformed uniform cartesian meshes. An equivalent number of uniform cartesian grids allow the comparison with the regular case, in two as well as in three dimensions.

- On the first neighborhood we use the least squares method, (see Proposition 2.4 above), as well as a method based on Green's Theorem and explained in Subsection 4.3.
- On the second neighborhood we test the least squares method and the second order gradient reconstruction explained in Subsection 4.4.

All test cases have the fixed advection velocities $\boldsymbol{c} = \frac{1}{\sqrt{8}}\left(-\sqrt{5}, \sqrt{3}\right)$ in two dimensions and $\boldsymbol{c} = \frac{1}{\sqrt{14}}\,(1, -3, 2)$ in three dimensions.

4.2 Definition of Curved Faces

In three dimensions, the deformed cartesian meshes contain curved faces as explained in Remark 2.1. In this case, the face $\mathcal{A}_{\alpha\beta}$ and its barycenter $\boldsymbol{x}_{\alpha\beta}$ are defined as follows. Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$ be the vertices shared by two adjacent cells $\mathcal{T}_\alpha$ and $\mathcal{T}_\beta$. Assume $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$ to be ordered so that the segments $\{\overline{\boldsymbol{v}_1\boldsymbol{v}_2}, \overline{\boldsymbol{v}_2\boldsymbol{v}_3}, \ldots, \overline{\boldsymbol{v}_l\boldsymbol{v}_1}\}$ form a closed path. Set $\boldsymbol{v}_{l+1} \triangleq \boldsymbol{v}_1$ and choose an arbitrary point $\boldsymbol{p}$. If the face $\mathcal{A}_{\alpha\beta}$ is defined as the union of the l triangles $\mathcal{A}_{\alpha\beta}^{(i)} = \overline{\boldsymbol{p}\boldsymbol{v}_i\boldsymbol{v}_{i+1}}$, then

$$|\mathcal{A}_{\alpha\beta}|\,(\boldsymbol{x}_{\alpha\beta} - \boldsymbol{p}) = \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{x} - \boldsymbol{p})\; dx \tag{4.1}$$

$$= \sum_{i=1}^{l} \int_{\mathcal{A}_{\alpha\beta}^{(i)}} (\boldsymbol{x} - \boldsymbol{p})\; dx = \sum_{i=1}^{l} \left|\mathcal{A}_{\alpha\beta}^{(i)}\right| \frac{1}{3}\left[(\boldsymbol{v}_i - \boldsymbol{p}) + (\boldsymbol{v}_{i+1} - \boldsymbol{p})\right].$$

Setting $\boldsymbol{x}_{\alpha\beta} = \boldsymbol{p}$ gives the implicit equation

$$\sum_{i=1}^{l} \left|\mathcal{A}_{\alpha\beta}^{(i)}\right| \frac{1}{3}\left[(\boldsymbol{v}_i - \boldsymbol{x}_{\alpha\beta}) + (\boldsymbol{v}_{i+1} - \boldsymbol{x}_{\alpha\beta})\right] = 0 \tag{4.2}$$

that can be solved for $\boldsymbol{x}_{\alpha\beta}$ by an iterative process. The resulting face $\mathcal{A}_{\alpha\beta}$ is the union of the l triangles $\mathcal{A}_{\alpha\beta}^{(i)} = \overline{\boldsymbol{x}_{\alpha\beta}\boldsymbol{v}_i\boldsymbol{v}_{i+1}}$. As explained in Remark 2.1, the face normal $\boldsymbol{a}_{\alpha\beta}$ is independent of the choice of $\boldsymbol{p}$.

4.3 Green Reconstruction

In this subsection, we describe a reconstruction method based on Green's Theorem, [38]. It is consistent for grids with curved faces. It requires the definition of the orthogonal projection $\boldsymbol{j}_{\alpha\beta}$ of $\boldsymbol{k}_{\alpha\beta}$ on $\boldsymbol{h}_{\alpha\beta}$, see Fig. 2.1. It is further useful to define the auxiliary point $\boldsymbol{y}_{\alpha\beta} = \boldsymbol{x}_\alpha + \boldsymbol{j}_{\alpha\beta}$. The idea is to consider a linear function $v\left(\boldsymbol{x}\right) = v_\alpha + \boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{x}_\alpha\right)$ and to derive a formula for its gradient $\boldsymbol{\sigma}$ by means of Green's Theorem

$$\int_{\mathcal{T}_\alpha} \boldsymbol{\nabla} v\left(\boldsymbol{x}\right)\, dx = \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\left(\boldsymbol{x}\right) v\left(\boldsymbol{x}\right)\, d\sigma \tag{4.3}$$

that v must satisfy. In this way, the resulting gradient formula is consistent by definition. The linearity of v implies $\boldsymbol{\nabla} v\left(\boldsymbol{x}\right) = \boldsymbol{\sigma}$, $v\left(\boldsymbol{x}\right) = v\left(\boldsymbol{y}_{\alpha\beta}\right) + \boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{y}_{\alpha\beta}\right)$ and $v_\alpha = v\left(\boldsymbol{x}_\alpha\right)$ where v_α is the mean value of v over the cell $\mathcal{T}_\alpha$. The value $v\left(\boldsymbol{y}_{\alpha\beta}\right)$ satisfies

$$v\left(\boldsymbol{y}_{\alpha\beta}\right) = v\left(\boldsymbol{x}_\alpha + \boldsymbol{j}_{\alpha\beta}\right) = v\left(\boldsymbol{x}_\alpha\right) + \frac{\left\|\boldsymbol{j}_{\alpha\beta}\right\|}{\left\|\boldsymbol{h}_{\alpha\beta}\right\|}\left(v\left(\boldsymbol{x}_\beta\right) - v\left(\boldsymbol{x}_\alpha\right)\right)\, . \tag{4.4}$$

Equation (4.3) becomes

$$\begin{aligned} \left|\mathcal{T}_\alpha\right| \boldsymbol{\sigma} = \sum_\beta \boldsymbol{a}_{\alpha\beta} v_\alpha + \sum_\beta \boldsymbol{a}_{\alpha\beta} \frac{\left\|\boldsymbol{j}_{\alpha\beta}\right\|}{\left\|\boldsymbol{h}_{\alpha\beta}\right\|}\left(v_\beta - v_\alpha\right) + \\ + \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\left(\boldsymbol{x}\right)\left[\boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{y}_{\alpha\beta}\right)\right] d\sigma\, . \end{aligned} \tag{4.5}$$

The first term on the right hand side of (4.5) is zero due to (2.3). The notation $\boldsymbol{y}_{\alpha\beta} - \boldsymbol{x}_\alpha = \boldsymbol{j}_{\alpha\beta}$ allows to write the identity

$$\begin{aligned} \left|\mathcal{T}_\alpha\right| \boldsymbol{\sigma} &= \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla}\left[\boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{x}_\alpha\right)\right] dx = \\ &= \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\left(\boldsymbol{x}\right)\left[\boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{y}_{\alpha\beta} + \boldsymbol{y}_{\alpha\beta} - \boldsymbol{x}_\alpha\right)\right] d\sigma \\ &= \sum_\beta \boldsymbol{a}_{\alpha\beta}\left(\boldsymbol{\sigma} \cdot \boldsymbol{j}_{\alpha\beta}\right) + \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\left(\boldsymbol{x}\right)\left[\boldsymbol{\sigma} \cdot \left(\boldsymbol{x} - \boldsymbol{y}_{\alpha\beta}\right)\right] d\sigma\, . \end{aligned} \tag{4.6}$$

Insertion of (4.6) in (4.5) and the collinearity of $\boldsymbol{j}_{\alpha\beta}$ and $\boldsymbol{h}_{\alpha\beta}$ give the linear relation between $\boldsymbol{\sigma}$ and the vector $\delta\mathfrak{v}$ with components $v_\beta - v_\alpha$

$$\sum_\beta \frac{\left\|\boldsymbol{j}_{\alpha\beta}\right\|}{\left\|\boldsymbol{h}_{\alpha\beta}\right\|}\left(\boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta}\right) \boldsymbol{\sigma} = \sum_\beta \frac{\left\|\boldsymbol{j}_{\alpha\beta}\right\|}{\left\|\boldsymbol{h}_{\alpha\beta}\right\|} \boldsymbol{a}_{\alpha\beta}\left(v_\beta - v_\alpha\right)\, . \tag{4.7}$$

The definition

$$\boldsymbol{n}'_{\alpha\beta} = \frac{\|\boldsymbol{j}_{\alpha\beta}\|}{\|\boldsymbol{h}_{\alpha\beta}\|}\boldsymbol{a}_{\alpha\beta} \tag{4.8}$$

simplifies (4.7) and results in

$$\sum_{\beta} \left(\boldsymbol{n}'_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta}\right) \boldsymbol{\sigma} = \sum_{\beta} \boldsymbol{n}'_{\alpha\beta} \left(v_\beta - v_\alpha\right) \ . \tag{4.9}$$

Finally, the definition of the matrix N'_α with row vectors $\boldsymbol{n}'_{\alpha\beta}$ allows to write the Green slope reconstruction formula (4.9) in matrix form as

$$\boldsymbol{\sigma} = S^{\mathrm{gr}}_\alpha \delta\mathfrak{v} = \left(N'^t_\alpha H_\alpha\right)^{-1} N'^t_\alpha \delta\mathfrak{v} \ . \tag{4.10}$$

4.4 Second Order Reconstruction

An alternative reconstruction method can be defined by the requirement that the gradient reconstruction be second order accurate. Let h denote the maximum cell diameter of the mesh and v be a sufficiently smooth function. A Taylor expansion of the mean value v_α of v on cell $\mathcal{T}_\alpha$ gives

$$\begin{aligned} v_\alpha = \frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} v\left(\boldsymbol{x}\right)\, dx = \\ \frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} \left[v\left(\boldsymbol{x}_\alpha\right) + \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_\alpha\right)^t \nabla^2 v\big|_{\boldsymbol{x}_\alpha}\left(\boldsymbol{x}-\boldsymbol{x}_\alpha\right) + O\left(h^3\right)\right] dx \end{aligned} \tag{4.11}$$

where $\nabla^2 v$ is the second derivative of v. In (4.11), the linear term in $\boldsymbol{x}-\boldsymbol{x}_\alpha$ is zero because $\boldsymbol{x}_\alpha$ is the barycenter of the cell $\mathcal{T}_\alpha$. With the definition of the symmetric matrix

$$X_\alpha = \frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} \left(\boldsymbol{x}-\boldsymbol{x}_\alpha\right) \otimes \left(\boldsymbol{x}-\boldsymbol{x}_\alpha\right)\, dx \tag{4.12}$$

the expansion (4.11) becomes

$$v_\alpha = \left[v\left(\boldsymbol{x}_\alpha\right) + \frac{1}{2}\mathrm{tr}\left(X^t_\alpha \ \nabla^2 v\big|_{\boldsymbol{x}_\alpha}\right) + O\left(h^3\right)\right] \tag{4.13}$$

Insertion of (4.13) in the general slope reconstruction (2.17)

$$\boldsymbol{\sigma}_\alpha = \sum_\beta \boldsymbol{s}_{\alpha\beta}\left(v_\beta - v_\alpha\right)$$

Table 4.1: Summary of Test Results

Dimension	Reconstruction	Stencil	Instability
2D	Least-Squares	First Neighborhood	No
2D	Green	First Neighborhood	Yes, but small
2D	Least-Squares	Second Neighborhood	No
2D	Second Order	Second Neighborhood	No
3D	Least-Squares	First Neighborhood	Yes
3D	Green	First Neighborhood	Yes
3D	Least-Squares	Second Neighborhood	No
3D	Second Order	Second Neighborhood	Yes

and expansion of $v\left(\boldsymbol{x}_\beta\right)$ at $\boldsymbol{x}_\alpha$ results in

$$\begin{aligned}\boldsymbol{\sigma}_\alpha = \sum_\beta \boldsymbol{s}_{\alpha\beta} \Bigg[& \nabla v|_{\boldsymbol{x}_\alpha} \boldsymbol{h}_{\alpha\beta} + \frac{1}{2} \boldsymbol{h}_{\alpha\beta}^t \, \nabla^2 v|_{\boldsymbol{x}_\alpha} \boldsymbol{h}_{\alpha\beta} \\ & + \frac{1}{2} \mathrm{tr}\left(X_\beta^t \, \nabla^2 v|_{\boldsymbol{x}_\alpha} \right) - \frac{1}{2} \mathrm{tr}\left(X_\alpha^t \, \nabla^2 v|_{\boldsymbol{x}_\alpha} \right) + O\left(h^3\right) \Bigg] \, .\end{aligned} \tag{4.14}$$

The second order accuracy requirement

$$\boldsymbol{\sigma}_\alpha = \nabla v|_{\boldsymbol{x}_\alpha} + O\left(h^2\right)$$

leads to the consistency condition (2.19) and to the second order accuracy condition

$$\sum_\beta \boldsymbol{s}_{\alpha\beta} \otimes \left[\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + X_\beta - X_\alpha \right] = 0 \, . \tag{4.15}$$

Let m_α be the size of the reconstruction stencil. Conditions (2.19) and (4.15) consist of d^2 and $\frac{1}{2} d^2 \left(d+1\right)$ linear equations, respectively. They can be solved for the $m_\alpha d$ reconstruction coefficients if $m_\alpha = \frac{1}{2} d \left(d+3\right)$. If m_α is larger, it is suitable to choose the minimum norm solution of systems (2.19) and (4.15).

4.5 Numerical Results

Table 4.1 gives a brief summary of the results. In the following, we examine point by point the observations and bring them into relation with the theoretical results of Subsection 3.6. Furthermore, we try to highlight the connection between the values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = \|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)}$ and the stability of the MUSCL scheme. Recall that the matrix $K_\alpha S_\alpha$ is invariant under scaling of the grid. This justifies the comparison of values of $K_\alpha S_\alpha$ across different grids.

Table 4.2: Least-Squares Reconstruction on the First Neighborhood in 2D : spectral abscissa ω_J and statistics of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

grid	spectral abscissa	average	maximum	90th percentile
triangular 1	-0.39404e-9	0.43621	0.54114	0.49491
triangular 2	0.23357e-9	0.42221	0.55833	0.45913
triangular 3	0.18482e-9	0.42372	0.59746	0.46769
triangular 4	0.39867e-10	0.41897	0.55139	0.44342
hybrid 1	0.23081e-9	0.42646	0.63910	0.52123
hybrid 2	0.63704e-10	0.41479	0.62273	0.49499
hybrid 3	-0.16952e-9	0.41004	0.61284	0.49313
hybrid 4	-0.32351e-10	0.40816	0.58758	0.47695
deformed cartesian 1	0.15821e-9	0.42557	0.65135	0.51010
deformed cartesian 2	0.52366e-10	0.43035	0.63335	0.51623
deformed cartesian 3	-0.21041e-9	0.43152	0.63201	0.51990
deformed cartesian 4	-0.32943e-9	0.43145	0.65605	0.51863

1. *First neighborhood reconstruction in two dimensions* : The numerical tests did not reveal any instabilities for the least squares reconstruction on irregular grids in two dimensions. The observed values for the norm $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ vary between 0.3 and 0.66. In the case of the least squares reconstruction, this appears to be sufficient to ensure stability for the chosen velocity direction. Table 4.2 displays the spectral abscissas and the most significant statistics of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ for this case. In contrast, the Green reconstruction of Section 4.3 produces a value of $\|R_\alpha\|_{\mathcal{L}(2,\infty)} \approx 1.5$ at some cells of the first hybrid grid. The resulting operator is still stable for the velocity $\boldsymbol{c} = \frac{1}{\sqrt{8}}\left(-\sqrt{5}, \sqrt{3}\right)$. However, for the velocity $\boldsymbol{c} = (1, 0)$, the operator becomes slightly unstable with a spectral abscissa $\omega_J \approx 0.0002$. The reason for this is that this grid presents distorted cells at the corners that create a problem for the Green reconstruction of Section 4.3. The least squares reconstruction produces much smaller values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)} \approx 0.64$ at these cells as predicted by Theorem 3.14. This particular example supports the conclusion of Section 3.7 that suggests that the least squares reconstruction is more likely to lead to robust schemes. The Green method is stable for the other grids that do not contain such distorted cells.
2. *Second neighborhood reconstruction in two dimensions* : Both the least squares and the second order reconstruction generate values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ that are smaller than those for reconstruction on the first neighborhood as suggested by Theorem 3.15. In the case of the least squares reconstruction, the values are about 50% smaller than those for the second order reconstruction in line with Theorem 3.14. The corresponding MUSCL operators are stable on all meshes.

Table 4.3: Least-Squares Reconstruction on the First Neighborhood in 3D: spectral abscissa ω_J and statistics of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

grid	spectral abscissa	average	maximum	90th percentile
tetrahedral 1	1.6539	0.57376	0.99051	0.67154
tetrahedral 2	-0.46968e-10	0.57143	1.0878	0.67217
tetrahedral 3	5.7716	0.56804	1.0533	0.65979
tetrahedral 4	7.5288	0.57435	1.0888	0.67144
hybrid 1	2.1612	0.54796	1.0820	0.65732
hybrid 2	5.5859	0.55320	1.0702	0.68159
hybrid 3	6.5645	0.53307	1.0962	0.66178
hybrid 4	7.2591	0.52921	1.1547	0.64271
deformed cartesian 1	-0.17017e-9	0.40784	0.54825	0.45403
deformed cartesian 2	0.42669e-10	0.41018	0.54821	0.45641
deformed cartesian 3	-0.63580e-10	0.41188	0.58191	0.46029
deformed cartesian 4	-0.55940e-10	0.41334	0.56309	0.46198

3. *First neighborhood reconstruction in three dimensions* : This is the case where both the least squares method and the Green method of Section 3.7 generate unstable schemes on tetrahedral and hybrid grids. Both the median and the average value of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ are larger than 0.5 and the maximum values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ are larger than 1. This shows that the first neighborhood is too small for slope reconstruction on such grids. Table 4.3 displays the spectral abscissas and the most significant statistics of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ for the least squares reconstruction. The tetrahedral grid 2 gives a stable operator but this is only valid for the chosen velocity direction. Fig. 4.1 shows the spectrum for the tetrahedral grid 3 with least squares reconstruction. At least two unstable modes are visible on the right of the imaginary axis.
4. *Second neighborhood reconstruction in three dimensions* : The least squares reconstruction leads to values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ that are much smaller than those observed for the first neighborhood reconstruction as predicted by Theorem 3.15. All instabilities disappear and this observation is a strong evidence for a link between the values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ and the appearance of unstable eigenmodes. On the other hand, the second order reconstruction results in an unstable operator for the second tetrahedral grid. This method produces comparatively large values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)} \geq 1.0$ at some cells. This is another example that suggests that the least squares reconstruction leads to more robust schemes.
5. *Cartesian meshes* : On uniform cartesian meshes $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = 0.35355$ for the first neighborhood reconstruction in two dimensions and three dimensions. The MUSCL operators on cartesian meshes are stable. It is important to note that these meshes have automatically larger stencil

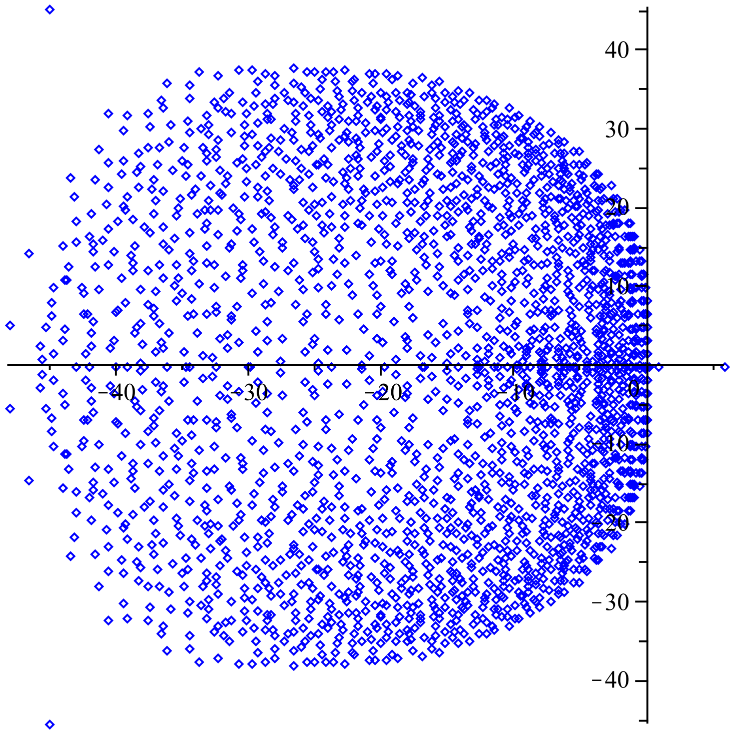


Figure 4.1: Unstable spectrum for a tetrahedral grid

sizes for the first order reconstruction. Again, this leads to smaller values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ as predicted by Theorem 3.15. This could explain the absence of unstable eigenmodes on such meshes.

6. *Strongly deformed cartesian meshes* : On these grids, each cell has the same number of neighbors as on a cartesian grid. The values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ are larger than those of cartesian grids but not enough to create instabilities. This is a hint that such meshes have a better stability behavior than tetrahedral and hybrid grids.
7. *Influence of the mesh type on stability* : The results of the tests seem to suggest that instabilities emerge only on tetrahedral and prismatic meshes for the first neighborhood reconstruction. Tetrahedra and prisms are the cells for which the size of the first neighborhood is smallest and the value of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ largest.
8. *Number of unstable modes* : We notice that the number of unstable eigenvalues seems to be very small, usually less than one percent of the total. This is of no help since rounding errors always introduce these modes into the numerical solution.

9. *Relationship between the local reconstruction map and eigenvalue stability* : The numerical evidence shows a strong correlation between the values of $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ and the existence of unstable eigenmodes. Unstable eigenmodes occur only on grids with cells where $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ approaches or exceeds 1. Unfortunately, we do not yet have a general theoretical proof for this relationship.

5 A three-dimensional CFD experiment

A simple three dimensional numerical experiment illustrates the interest of the preceding stability analysis for compressible fluid dynamics. The purpose of the experiment is to underline the following points:

1. The instabilities observed for the linear advection equation in Sec. 4.5 arise also in simulations of the nonlinear equations of compressible fluid dynamics, i.e. the Navier-Stokes or the inviscid Euler equations.
2. The origin of this specific type of instability is a spatial discretization that is unstable in the sense of definition 3.2, meaning that these instabilities are a linear phenomenon and not genuinely nonlinear.
3. In the presence of such instabilities, the slope limiters are necessary not only for the monotonicity of the MUSCL scheme but also to suppress these linear instabilities. The result of the computation depends in this case very much on the properties of the specific slope limiter.

In order to sustain these statements, it is useful to examine a typical situation where slope limiters should not be necessary for the stability of the scheme but should only serve to avoid spurious oscillations. A classical and very simple example for such a flow is the advection of a contact discontinuity in a three dimensional channel. The experiment has been performed with the package CEDRE developed at ONERA for applications in aerothermochemistry.

The computational domain is a rectangular channel of square section, given by

$$\Omega = \left\{ (x,y,z) \in \mathbb{R}^3 \,|\, -1 \le x \le 5 \,,\, 0 \le y \le 0.1 \,,\, 0 \le z \le 0.1 \right\} .$$

The initial condition is a contact discontinuity located at $x = 0$ in the channel. The discontinuity consists only of a jump in temperature and mass density, whereas pressure and velocity are uniform. The mesh, shown in fig. 5.1 , is made of $N \approx 900$ tetrahedra. The discontinuity evolves along the positive x axis at a velocity of $100\,\frac{\mathrm{m}}{\mathrm{s}}$. The boundary conditions consist of a subsonic inflow at $x = -1$, a subsonic outflow at $x = 5$ and solid walls.

The conserved quantities mass density, momentum and energy are denoted by

$$q = (\rho, \rho v_x, \rho v_y, \rho v_z, \rho E)$$

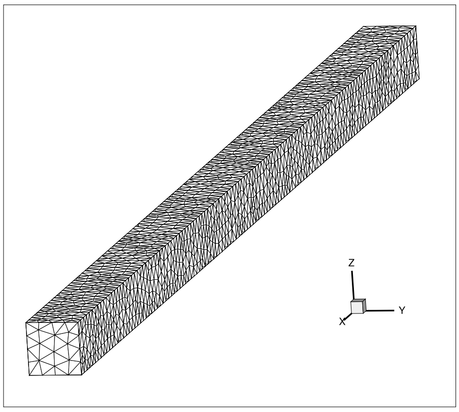


Figure 5.1: Tetrahedral grid of a channel; the flow is along direction x

and evolve according to the Euler equations for an ideal compressible gas. The computation uses a numerical flux of the Roe type and an explicit third order Runge-Kutta time-stepping scheme. The time step is $\Delta t = 10^{-5}$ and the CFL number is 0.05.

The propagation of a contact discontinuity at constant velocity behaves in a manner very similar to a solution of the linear advection equation. As no shock is involved, the numerical scheme without slope limiters should at least be stable, even if the lack of monotonicity generates spurious oscillations.

The numerical experiment allows to compare two reconstruction methods :

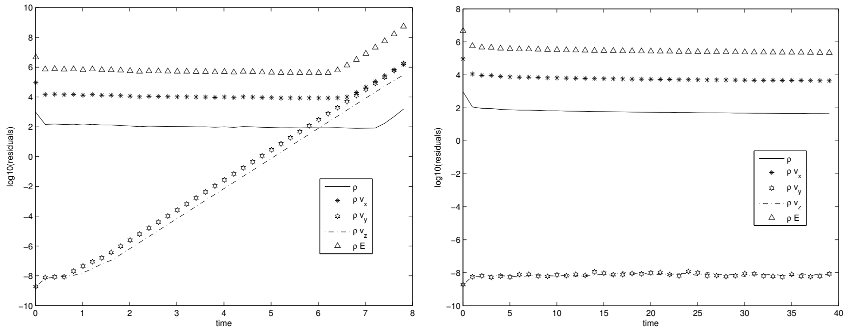
1. The least squares gradient reconstruction on the first neighborhood. According to the experiments of Sec. 4.5, this method leads to an unstable MUSCL discretization of the linear advection equation on tetrahedral grids.
2. A gradient reconstruction on the second neighborhood that gives a stable MUSCL discretization on tetrahedral grids.

Fig. 5.2a and 5.2b display the history of the residuals

$$t \mapsto \max_{\Omega} \left| \frac{dq}{dt} \right|$$

using a linear scale for the time and a logarithmic scale for the residuals. Fig. 5.3a and 5.3b present the history of the different components of the velocity.

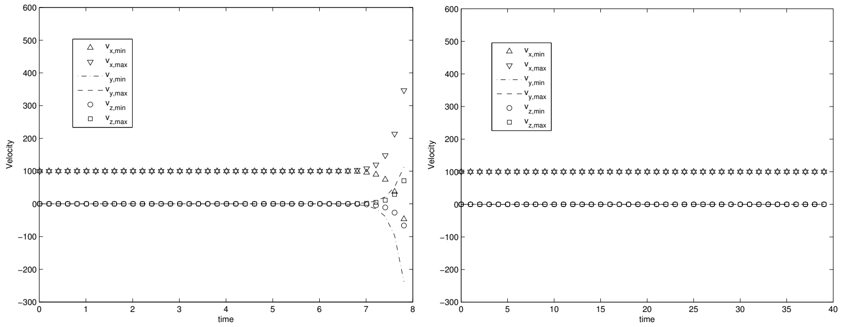
The results for the unstable first neighborhood reconstruction show that the residuals of ρv_y and ρv_z which are initially very small grow to reach



(a) 800 time iterations, first neighborhood interpolation of the gradient

(b) 4000 time iterations, second neighborhood interpolation of the gradient

Figure 5.2: History of the residuals of $\frac{d}{dt}[\rho, \rho u_x, \rho u_y, \rho u_z, \rho E]$ without slope limiters



(a) 800 time iterations, first neighborhood interpolation of the gradient

(b) 4000 time iterations, second neighborhood interpolation of the gradient

Figure 5.3: History of the minimum and maximum of the velocity without slope limiters

the magnitude of the residual of ρv_x after a few hundred time steps. The fact that the curves of the residuals of ρv_y and ρv_z are straight lines indicates that the growth is exponential. This is also visible on the plot of the components of the velocity in Fig. 5.3a. When the stable discretization is used, the residuals as well as the minima and maxima of the velocity stay constant although no slope limiter is used. On cartesian meshes, both discretization methods yield a stable solution of this specific problem.

This suggests that the specific type of instability visible in Fig. 5.2a is caused by an unstable eigenmode of the linear advection equation. At $t = 0$, ρv_y and ρv_z are not completely zero due to rounding errors. As these components are advected along the x-direction at constant velocity v_x, the unstable eigenmode begins to grow exponentially. In the case of the second neighborhood reconstruction, there is no instability and the computation can be carried out without any kind of nonlinear limitation procedure, see Fig. 5.2b. The same holds true for structured cartesian grids.

Finally, the computation has been repeated with different slope limiters available in CEDRE. A recently implemented slope limiter that is based on the maximum principle of Barth [3,4] is able to contain the linear instability and to restore the correct solution. However, this is not the case for other slope limiters that are based on empirical formulas.

A good example for the usefulness of the preceding analysis is the large eddy simulation of a hot supersonic jet that has been the subject of an experimental study by Seiner and Ponton [45]. Simulations with CEDRE on tetrahedral grids were not able to capture the transition of the jet to turbulent flow, whereas simulations on hexahedral grids delivered numerical results in accordance with the experimental data. However, it was previously impossible to determine if the problem is caused by the slope limiters or the gradient reconstruction because the simulation could not be carried out without slope limiters on tetrahedral grids.

With the stable gradient reconstruction on the second neighborhood, the computation can be performed with slope limiters active only inside the nozzle. In this particular setting, the jet becomes turbulent on tetrahedral grids in the same way as on hexahedral grids. This shows clearly that the slope limiters are responsible for preventing the transition to turbulence. In a second step, the result of this particular computation has been used as a turbulent initial condition for a computation of the jet with the standard MUSCL discretization using slope limiters. This turbulent initial condition gives much better results than the non turbulent initial condition. Details can be found in [24].

6 Conclusion

The subject of this article is the eigenvalue stability of the semi-discrete equation obtained by a MUSCL discretization of linear advection without slope limiters. The main purpose of the study is to explore the influence of the mesh type, the slope reconstruction method and the stencil size on stability. To isolate the influence of these factors, the stability of the scheme has been studied without slope limiters.

A general result proves the stability of the first order finite volume discretization of linear advection on arbitrary meshes. For the MUSCL scheme, the analysis proceeds in two distinct steps.

1. The paper introduces a new criterion to evaluate the impact of the piecewise linear slope reconstruction on the stability of the MUSCL scheme. This criterion defines an *approximate and qualitative measure* to identify the reconstruction methods that are best suited to give a stable

MUSCL operator (3.7). Such a criterion is indispensable because the exact relationship between the eigenvalues of the MUSCL operator and the slope reconstruction is too difficult to analyze on general unstructured grids. The criterion uses a local property in each cell, the local reconstruction map given by Definition 3.7. This local reconstruction map is a dimensionless matrix that is invariant under scalings of the grid. If certain norms of this matrix are smaller for one particular reconstruction method, then there are good reasons to think that this method leads to a more robust scheme.

2. The second step consists of looking for an exact minimizer of this criterion. For a certain family of norms, it is possible to identify the least squares reconstruction as a minimizer as shown in Theorem 3.14. Furthermore, it is possible to show that an extension of the reconstruction stencil cannot increase the value of the criterion. Under a simple rank condition, the extension of the stencil leads even to a strictly decreasing value of the criterion.

The subsequent analysis and the theorems of Subsection 3.6 provide two practical conclusions for the choice and design of reconstruction methods and their stencils.

1. Theorem 3.14 shows that the least squares slope reconstruction is a minimizer of the new criterion of Definition 3.11 in a certain family of norms. This result suggests that if the least squares slope reconstruction gives an unstable scheme, then any consistent slope reconstruction is also very likely to lead to an unstable scheme. This result can be loosely interpreted as a result of optimality of the least squares reconstruction but this interpretation is of course not completely rigorous. This conclusion is supported by the fact that the Green reconstruction of Section 4.3 had to be modified in CEDRE because it produced unusually large gradients on distorted meshes.
2. The result of Theorem 3.15 suggests that a larger reconstruction stencil should lead to a more robust scheme, at least for the least squares reconstruction. In section 4, this hypothesis has been tested numerically and it turns out to be particularly true for three-dimensional meshes. This could explain why hexahedral grids have better stability properties than tetrahedral grids.

Numerical computations of the spectra of MUSCL discretization operators complete and confirm the theoretical part of this paper. They show a strong correlation between a local property of the slope reconstruction, the so called local reconstruction map, and the appearance of unstable eigenmodes. The discretization seems to be stable on all two-dimensional meshes. Significant instabilities arise on tetrahedral grids when the reconstruction stencil is the first neighborhood of the cell, showing that this stencil is too small for this type of mesh. The instabilities disappear when the least squares reconstruction is used on a larger stencil, in this case the second neighborhood. However, they can emerge again if an alternative reconstruction method is used, in this example the second order gradient reconstruction. These numerical results underpin the theoretical conclusions.

The results have allowed to design a new consistent slope reconstruction method in CEDRE that is based on the second neighborhood and gives a stable discretization of the linear advection equation, [25,24]. With this new reconstruction, the computation of a subsonic flow over a deep cavity can now be computed without slope limiters. This situation allows to measure the impact of different slope limiters on the solution [24]. In the case of the hot supersonic jet, it was possible to carry out the computation with slope limiters only active inside the nozzle. The result establishes clearly that the dampening of the jet noise is caused by the slope limiters and not by the piecewise linear reconstruction [24].

References

1. Abgrall, R.: On Essentially Non-oscillatory Schemes on Unstructured Meshes : Analysis and Implementation. J. Comput. Phys. **114** (1) : 45-58, (1994).
2. Agarwal R.K., Halt D.W.: A Compact High-order Unstructured Grids Method for the Solution of Euler Equations. Int. J. Numer. Meth. Fluids **31** : 121-147 (1999).
3. Barth, T., Ohlberger, M.: Finite Volume Methods: Foundation and Analysis. In: Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes, John Wiley and Sons (2004).
4. Barth, T.: Numerical Methods for Conservation Laws on Structured and Unstructured Meshes. VKI Lecture Series (2003).
5. Ben-Artzi, M., Falcovitz J.: *Generalized Riemann Problems in Computational Fluid Dynamics* (2003) Cambridge Univ. Press.
6. Berger M., Aftosmis M.J., Murman S.M.: Analysis of Slope Limiters on Unstructured Grids. NAS Technical Report NAS-05-007, 43th AIAA (2005).
7. Berthon, C.: Why the MUSCL-Hancock scheme is L^1-stable. Numer. Math. **104**, 27-46, (2006).
8. Berthon, C.: Stability of the MUSCL schemes for the Euler equations. Comm. Math. Sci. **3** 133-157, (2005).
9. Bertier, N.: Simulation des grandes échelles en aérothermique sur des maillages non structurés généraux. Thesis, Univ. Paris 6 (2006).
10. Bouchut, F., Perthame, B., Bourdarias, C.: A MUSCL method satisfying all the numerical entropy inequalities. Math. of Comp., **65** 1439-1461, (1996).
11. Buffard, T., Clain, S. : Multi-Slope MUSCL Methods for Unstructured Meshes, preprint Univ. Clermont-Ferrand (2006).
12. Chainais-Hillairet, C. : Second Order Finite Volume Schemes for a Nonlinear Hyperbolic Equation : Error Estimate. Math. Meth. Appl. Sci. **23** (5) : 467-490 (2000).
13. Chakravarthy, S., Harten, A.: Multi-dimensional ENO Schemes for General Geometries. NASA Contractor Report NASA-CR-187637, ICASE Report No. 91-76 (1991).
14. Chevalier P., Courbet B., Dutoya D., Klotz P., Ruiz E., Troyes J., Villedieu P.: CEDRE: Development and Validation of a Multiphysic Computational Software. First European Conference for Aerospace Sciences (EUCASS), Moscou, Russie (2005).

15. Colella, P., and Woodward, P.: The numerical simulation of two-dimensional flows with strong shocks, J. Comput. Phys. **54**, 115-173 (1984).
16. Camarri S., Salvetti M.V., Koobus B., Dervieux A.: A Low-diffusion MUSCL Scheme for LES on Unstructured Grids. Comput. Fluids **33 (**9) : 1101-1129 (2004).
17. Després, B.: An Explicit A Priori Estimate for a Finite Volume Approximation of Linear Advection on Non-Cartesian Grids. SIAM J. Numer. Anal. **42** (2) : 484-504 (2004).
18. Deuflhard, P., Bornemann F.: *Scientific Computing with Ordinary Differential Equations*. Springer Verlag, Berlin (2002).
19. Eymard, R., Gallouet, T., Herbin, R. : *Finite Volume Methods, Handbook of Numerical Analysis*, vol. 5, Ciarlet, Lions ed., North-Holland (1997).
20. Godlewski, E., and Raviart, P.A.: *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. (2002), Springer Verlag.
21. Golub, G.H., Van Loan, C.F.: *Matrix Computations* (3rd ed.). Johns Hopkins University Press (1996).
22. Gottlieb, S., Shu, C.W., Tadmor, E.: Strong Stability-preserving High-order Time Discretization Methods. SIAM Rev. **43** (1): 89-112 (2001).
23. Gustafsson B.: The Godunov-Ryabenkii condition: The beginning of a new stability analysis, Tech. Report 1999-14, Uppsala Univ., (1999).
24. Haider, F.: Discrétisation spatiale en maillage non-structuré de type général et applications LES. Thesis, Univ. Paris 6 (2009).
25. Haider, F., Croisille, J-P., Courbet, B.: Stability of the MUSCL method on general unstructured grids for applications to compressible flows. 5th ICCFD, July 2008, Séoul.
26. Harten, A.: High resolution schemes for hyperbolic conservation laws. J. Comput. Phys. **49**, 357-393, (1983).
27. Horn, R.A., and Johnson, C.R.: *Matrix Analysis*. Cambridge University Press (1985).
28. Horn, R.A., and Johnson, C.R.: *Topics in Matrix Analysis*. Cambridge University Press (1991).
29. Hirsch, C..: *Numerical Computation of Internal and External Flows*, vol. 1 & 2, John Wiley & Sons (2001).
30. Householder, A.S.: *The Theory of Matrices in Numerical Analysis*. Dover Publications, New York (1975).
31. Hubbard M.E., Multidimensional Slope Limiters for MUSCL-type Finite Volume Schemes on Unstructured Grids. J. Comput. Phys. **155** (1) : 54-74 (1999).
32. Iske, A., Sonar, Th.: On the Structure of Function Spaces in Optimal Recovery of Point Functionals for ENO-Schemes by Radial Basis Functions. Numer. Math. **74** : 177-201 (1996).
33. Jiang, G.S. , Shu, C.W.: Efficient Implementation of Weighted ENO Schemes. J. Comput. Phys. **126**(1) : 202-228 (1996).
34. Kröner, D.: *Numerical Schemes for Conservation Laws*. John Wiley and Sons, Chichester, and B.G. Teubner, Stuttgart (1997).
35. Khosla S., Dionne P.J., Lee M.E., Smith C. E..: Using Fourth Order Accurate Spatial Integration on Unstructured Meshes to Reduce LES Run Times. AIAA 2008-782, 46th AIAA (2008).
36. Kyu Hong Kim, Chongam Kim : Accurate, Efficient and Monotonic Numerical Methods for Multi-dimensional Compressible Flows Part II : Multi-dimensional Limiting Process. J. Comput. Phys. **208** (2) : 570-615 (2005).
37. Larchevêque L., Sagaut P., Mary, I., Labbé, O.: Large-eddy Simulation of a Compressible Flow Past à Deep Cavity. Phys. Fluid **15**(1) : 193-210 (2003).
38. Leterrier, N.: Discrétisation spatiale en maillage non-structuré de type général. Thesis, Univ. Paris 6 (2003).

39. Leveque, R..: *Finite Volume Methods for Hyperbolic Problems*. (2002), Cambridge Univ. Press.
40. Lupoglazoff N., Rahier G., Vuillot F. Application of the CEDRE Unstructured Flow Solver to Jet Noise Computations. First European Conference for Aerospace Sciences (EUCASS), Moscou, Russie, 2005.
41. Merlet, B., Vovelle, J.: Error estimates for the finite volume scheme applied to advection equation. Numer. Math. **106**, 129-155, (2007).
42. Perthame, B., Qiu, Y.: A variant of Van Leer's method for multidimensional systems of conservation laws. Jour Comp. Physics, **112**, 370-381, (1996).
43. Reddy, S.C., Trefethen, N.: Stability of method of lines. Numer. Math., **62**, 235-267, (1992).
44. Reichstadt S., Bertier, N., Ristori, A., and Bruel, P.: Towards LES of Mixing Processes inside a Research Ramjet Combustor. 18th Symposium ISOABE, ISABE Paper-2007-1188, Beijing, China (2007).
45. Seiner, J.M.; Ponton, M.; Jansen, B.J. & Lagen, N.T.: The effects of temperature on supersonic jet noise emission. DGLR/AIAA, 92-02-046: 295-307 (1992).
46. Selva, G.: Méthodes itératives pour l'intégration implicite des équations de l'aérothermochimie sur des maillages non-structurés. Thesis, Ecole Centrale de Paris (1998).
47. Shu, C.W., Osher, S.: Efficent Implementation of Essentially Non-oscillatory Shock-capturing Schemes. J. Comput. Phys. **77** (2): 439-471 (1988).
48. Shu, C.W.: High Order ENO and WENO Schemes. In: High-Order Methods for Computational Physics. Lecture Notes in Computational Science and Engineering (Vol. 9). Edited by Barth, T.J. and Deconinck, H., Springer Verlag Heidelberg (1999).
49. Sonar, Th.: Optimal Recovery Using Thin Plate Splines in Finite Volume Methods for the Numerical Solution of Hyperbolic Conservation Laws. IMA J. Num. Anal. **16** (4), 549-581 (1996).
50. Strikwerda J.: *Finite Difference Schemes and Partial Differential Equations*, (1989) Wadsworth and Brooks/Cole Publ.
51. Thornber, B., Mosedale, A., Drikakis, D., Youngs, D., Williams, R.J.R. : An Improved Reconstruction Method for Compressible Flows with Low Mach Number Features. J. Comput. Phys. **227** (10) : 4873-4894 (2008).
52. Toro E.: *Riemann solvers and numerical methods for fluid flows*, (1999), Springer-verlag, Berlin.
53. Van Leer, B.: Towards the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection, J. Comput. Phys. **23**, 276-299 (1977).
54. Van Leer, B.: Towards the Ultimate Conservative Difference Scheme. V. A Second Order Sequel to Godunov's method, J. Comput. Phys. **32**, 101-136 (1979).
55. Varga, R.S.: *Geršgorin and His Circles*. Springer Verlag, Berlin (2004).