Journal of Computational Physics 229 (2010) 978–1016

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number

Stéphane Dellacherie *

Commissariat à l'Énergie Atomique, CEA,DEN,DM2S,SFME, Laboratoire d'Études Thermohydrauliques des Réacteurs, CEA – Centre de Saclay, F-91191 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 17 November 2008
Received in revised form 25 September 2009
Accepted 29 September 2009
Available online 9 October 2009

Keywords:
Compressible Euler system
Low Mach number flow
Roe scheme
Roe–Turkel scheme
AUSM$^+$-up scheme
Linear wave equation
Hodge decomposition

ABSTRACT

We propose a theoretical framework to clearly explain the inaccuracy of Godunov type schemes applied to the compressible Euler system at low Mach number on a Cartesian mesh. In particular, we clearly explain why this inaccuracy problem concerns the 2D or 3D geometry and does not concern the 1D geometry. The theoretical arguments are based on the Hodge decomposition, on the fact that an appropriate well-prepared subspace is invariant for the linear wave equation and on the notion of first-order modified equation. This theoretical approach allows to propose a simple modification that can be applied to any colocated scheme of Godunov type or not in order to define a large class of colocated schemes accurate at low Mach number on any mesh. It also allows to justify colocated schemes that are accurate at low Mach number as, for example, the Roe–Turkel and the AUSM$^+$-up schemes, and to find a link with a colocated incompressible scheme stabilized with a Brezzi–Pitkäranta type stabilization. Numerical results justify the theoretical arguments proposed in this paper.

1. Introduction

It is sometimes essential to take into account compressibility phenomena even when a flow is at low Mach number. Moreover, a flow may also be at low Mach number only in a part of the physical domain. In such situations, it is important to model the flow with the compressible Navier–Stokes system.

Since the compressible Euler system is the compressible Navier–Stokes system without the physical diffusive terms, any study at low Mach number may firstly concern the compressible Euler system. Conservative finite volume schemes as Godunov type schemes [1–3] are colocated schemes that are well adapted to capture shock waves solution of the compressible Euler system. Nevertheless, it is well known that Godunov type schemes are not accurate at low Mach number [4–10]. This inaccuracy is characterized by the creation of spurious pressure and velocity waves that avoid the velocity field to be close to a divergence-free field. Let us note that this inaccuracy concerns other colocated compressible schemes [11,12]. This inaccuracy problem comes from a loss of information between the continuous and discrete levels. Indeed, at the continuous level, the solution of the compressible Euler system converges toward the solution of the incompressible Euler system when the Mach number goes to zero. Nevertheless, many numerical experiments show that at low Mach number, the numerical solution given by a colocated compressible scheme may be far from an incompressible numerical solution. In [6,10], the origin of the inaccuracy is explained by studying the 1D-formulation of Godunov type schemes through a formal asymptotic development based on the Mach number. Such studies allow, firstly, to partly find the origin of the inaccuracy and, secondly, to

* Tel.: +33 1 69 08 15 11; fax: +33 1 69 08 10 87.
E-mail address: stephane.dellacherie@cea.fr.

0021-9991/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2009.09.044

propose modifications of Godunov type schemes to increase the accuracy at low Mach number. For example, in [10], the Roe scheme with pressure correction is proposed. This scheme consists in modifying the numerical viscosity by adding a pressure correction on the momentum equation. The Roe–Turkel [4,5,7] and VFRoe–Turkel [8,9] schemes are other modified Godunov type schemes; the schemes proposed in [6,13] are similar to the Roe–Turkel scheme. All these schemes are built by modifying the numerical diffusion with the Turkel preconditioning matrix [14,15]. In [7–9], the Roe–Turkel and VFRoe–Turkel schemes are justified with a formal asymptotic development based on the Mach number; *idem* in [6]. Let us underline that the Roe–Turkel and VFRoe–Turkel schemes are much more complicated to implement than the Roe scheme with pressure correction proposed in [10]. All these modified Godunov type schemes give good numerical results. In [12], Liou proposes a flux splitting type scheme – named AUSM$^+$-up scheme – that is also accurate at low Mach number. In [16–18], other colocated schemes accurate at low Mach number are proposed.

We propose in this work a theoretical framework to clearly explain the inaccuracy of Godunov type schemes at low Mach number when the mesh is Cartesian. This theoretical framework is based on a Schochet's result [19], on the Hodge decomposition, on the fact that an appropriate well-prepared subspace is invariant for the linear wave equation and on the notion of first-order modified equation. In particular, we show that the inaccuracy problem does not exist in 1D geometry that is to say only exists in 2D or 3D geometry. Let us note that in [8], the linear wave equation is also studied by Guillard et al. But, our analysis seems to be more direct and, especially, allows to clearly identify the invariance property that is the key argument. Our analysis underlines also that the inaccuracy of Godunov type schemes at low Mach number can be explained (at least partly) with simple linear arguments by only analyzing the linear wave equation. Nevertheless, it is important to note that the proposed theoretical approach would not have been possible without the previous works of Schochet and Guillard et al. We also propose a formal approach that is coherent with our theoretical results and we compare this formal approach to other formal approaches [4,6,8,10] that seem to be less precise. The proposed theoretical approach allows us to conjecture the existence of a large class of colocated schemes that are accurate at low Mach number on any mesh. This class is named *low Mach X schemes* and consists in a simple modification of any X scheme of Godunov type ($X \in \{Roe, VFRoe, \ldots\}$) or not ($X$ = kinetic scheme [20] for example). Numerical results show that the *low Mach Roe scheme* and the *low Mach VFRoe scheme* (i.e. $X = Roe$ and $X = VFRoe$) are accurate at low Mach number. Let us underline that the proposed theoretical approach justifies the Roe–Turkel [4,5,7] (or VFRoe–Turkel [8,9]) scheme and the Roe scheme with pressure correction [10] since, in the linear case, they are respectively similar and identical to the *low Mach Godunov scheme*. Moreover, it is also possible to prove that the schemes proposed in [12,16–18] are similar or identical to a *low Mach X scheme* when the Mach number goes to zero.

The outline of this paper is the following. In Section 2, we recall the derivation of the incompressible Euler system from the compressible Euler system with formal arguments and with theoretical arguments due to Schochet [19]. Then, we show how to prove the Schochet's result in the linear case with simple arguments. This Section 2 allows us to introduce the notion of well-prepared subspace and a sufficient condition allowing to avoid the creation of spurious pressure and velocity waves. In Section 3, we propose a simple formal approach showing that the inaccuracy problem only exists in 2D or 3D geometry. In Section 4, we explain with theoretical arguments why Godunov type schemes are inaccurate only in 2D or 3D geometry. In Section 5, we propose to apply a simple modification to any colocated X scheme in order to define a large class of *low Mach X schemes* accurate at low Mach number. This modification is justified in the case of the linear Godunov scheme. Then, we show that Roe–Turkel type schemes [7–9] and Godunov type schemes with pressure correction [10] are respectively similar and identical to the *low Mach Godunov scheme* in the linear case. We also underline that the AUSM$^+$-up scheme [12] and the colocated schemes proposed in [16–18] are respectively similar and identical to a *low Mach X scheme* when the Mach number goes to zero. Moreover, we formally show that there exists a link between the proposed *low Mach Godunov scheme* and a colocated incompressible scheme stabilized with a Brezzi–Pitkäranta type stabilization [21]. In Section 6, we introduce a more general theoretical framework allowing to propose other colocated schemes that do not create spurious pressure and velocity waves. Finally, we show in Section 7 numerical results that justify the theoretical arguments proposed in this paper.

2. Convergence of a compressible flow toward an incompressible flow

We recall the classical formal derivation of the incompressible Euler system from the compressible Euler system [22]. This recall will allow us to formally introduce the notion of well-prepared initial condition. This will be essential in the sequel. Then, we recall the theoretical derivation proposed in [19] and we prove the result in the linear case with simple arguments. This linear study will allow us to, firstly, define in Section 2.4 a sufficient condition to avoid the creation of spurious (pressure and velocity[1]) waves, to, secondly, explain in Section 4 the inaccuracy of Godunov type schemes at low Mach number and to, thirdly, propose in Section 5 a simple curative solution. In Section 6, we will propose a weaker sufficient condition to avoid the creation of spurious waves.

2.1. Formal derivation

The dimensionless compressible Euler system is given by

[1] In the sequel, we write *spurious waves* for *spurious pressure and velocity waves*.

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho u) = 0, & \text{(a)} \\ \partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) + \dfrac{\nabla p}{M^2} = 0, & \text{(b)} \\ \partial_t (\rho E) + \nabla \cdot [(\rho E + p) u] = 0, & \text{(c)} \\ E = \frac{M^2 u^2}{2} + \varepsilon(\rho, p) & \text{(d)} \end{cases} \tag{1}$$

where M is the Mach number ($M \ll 1$ at low Mach number). In (1), ρ, u, p, ε and E are respectively the density, the velocity, the pressure, the internal energy and the total energy. System (1) is closed as soon as the equation of state $\varepsilon(\rho, p)$ (or $p(\rho, \varepsilon)$) is known. System (1) is equivalent to the non-conservative system

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho u) = 0, & \text{(a)} \\ \rho [\partial_t u + (u \cdot \nabla) u] + \dfrac{\nabla p}{M^2} = 0, & \text{(b)} \\ \partial_t p + u \cdot \nabla p + \rho a^2(\rho, p) \nabla \cdot u = 0 & \text{(c)} \end{cases} \tag{2}$$

for any regular solution. In (2), $a(\rho, p) = \sqrt{\frac{\partial p}{\partial \rho}(\rho, \varepsilon) + \frac{p}{\rho^2} \frac{\partial p}{\partial \varepsilon}(\rho, \varepsilon)}$ is the sound velocity (we admit that $\varepsilon(\rho, p)$ is such that $a(\rho, p) > 0$ for any $(\rho, p) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+$). By supposing that

$$\begin{cases} p(t,x) = p_0(t,x) + M p_1(t,x) + M^2 p_2(t,x) + \cdots, & \text{(a)} \\ u(t,x) = u_0(t,x) + M u_1(t,x) + \cdots, & \text{(b)} \end{cases} \tag{3}$$

by injecting (3) in (2) and by separating the $\mathcal{O}(M^{-2})$ and $\mathcal{O}(M^{-1})$ terms,[2] we formally find that $\nabla p_0 = 0$ and $\nabla p_1 = 0$. Thus, we can replace (3)(a) with $p(t,x) = p_0(t) + M^2 p_2(t,x) + \cdots$. Then, by taking the $\mathcal{O}(1)$ terms, we obtain

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho u_0) = 0, & \text{(a)} \\ \rho [\partial_t u_0 + (u_0 \cdot \nabla) u_0] + \nabla p_2 = 0, & \text{(b)} \\ \nabla \cdot u_0 = -\dfrac{\dfrac{d}{dt} p_0(t)}{\rho a^2(\rho, p_0)}. & \text{(c)} \end{cases} \tag{4}$$

When we solve system (1) on a bounded domain Ω with a slipping or periodic boundary condition, we have $\int_\Omega \nabla \cdot u_0 dx = 0$. Thus, by integrating (4)(c) on Ω, we find that $\frac{d}{dt} p_0(t) = 0$ since $\rho a^2(\rho, p_0) > 0$. This means that the pressure is given by $p(t,x) = p_* + M^2 p_2(t,x) + \cdots$ where p_* is a constant of order 1. Coming back to (4)(c), we obtain that $\nabla \cdot u_0 = 0$. Coupled to (4)(a), this gives $\partial_t \rho + u_0 \cdot \nabla \rho = 0$. By imposing $\rho(t = 0, x) = \rho_*(x)$ where ρ_* is a strictly positive function such that $\mathcal{O}(\|\rho_*\|) = 1$ ($\|\cdot\|$ is an appropriate norm), we finally "obtain" that, at low Mach number (i.e. $M \ll 1$) and when the initial conditions are *well-prepared* in the sense

$$\begin{cases} p(t=0,x) = p_* + \mathcal{O}(M^2)(x), \\ u(t=0,x) = \hat{u}(x) + \mathcal{O}(M)(x) \end{cases} \quad \text{with} \quad \begin{cases} p_* = C_p^{st}, \\ \nabla \cdot \hat{u} = 0 \end{cases} \tag{5}$$

(C_p^{st} is a strictly positive constant of order 1; $\|\hat{u}\|$ is also of order 1), the solution (ρ, u, p) of compressible Euler system (1) is close to the solution (ρ, u, p) of the incompressible Euler system with variable density

$$\begin{cases} \partial_t \rho + u \cdot \nabla \rho = 0, \\ \rho(t=0,x) = \rho_*(x), \\ \nabla \cdot u = 0, \\ \partial_t u + (u \cdot \nabla) u + \dfrac{\nabla \Pi}{\rho_*} = 0, \\ u(t=0,x) = \widehat{u}(x), \\ p = p_*. \end{cases}$$

2.2. Theoretical derivation in the non-linear barotropic case

We now recall the convergence result proposed by Schochet [19] in the barotropic case. The dimensionless barotropic Euler system is given by

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) + \dfrac{\nabla p(\rho)}{M^2} = 0. \end{cases} \tag{6}$$

[2] In all the paper, the notation $\mathcal{O}(\alpha)$ means *of the order of* α.

In (6), we suppose that $p(\rho)$ is a strictly increasing function from $\mathbb{R}_*^+$ to $\mathbb{R}_*^+$. Thus, the sound velocity $a(\rho) = \sqrt{p'(\rho)}$ is strictly positive. By defining the variable r with

$$\rho(t,x) := \rho_* \left[1 + \frac{M}{a_*} r(t,x)\right] \tag{7}$$

($\rho_* = C_\rho^{st}$ where C_ρ^{st} is a strictly positive constant of order 1 and $a_* = \sqrt{p'(\rho_*)}$), for any regular solution, system (6) is equivalent to

$$\partial_t q + \mathcal{H}(q) + \frac{\mathcal{L}}{M}(q) = 0 \tag{8}$$

where

$$\begin{cases} q = \begin{pmatrix} r \\ u \end{pmatrix}, & \text{(a)} \\ \mathcal{H}(q) = \begin{pmatrix} u \cdot \nabla r \\ (u \cdot \nabla) u \end{pmatrix} := (u \cdot \nabla) q, & \text{(b)} \\ \mathcal{L}(q) = \begin{pmatrix} (a_* + Mr) \nabla \cdot u \\ \dfrac{p'\left[\rho_*\left(1 + \frac{M}{a_*} r\right)\right]}{a_*\left(1 + \frac{M}{a_*} r\right)} \nabla r \end{pmatrix}. & \text{(c)} \end{cases} \tag{9}$$

The operators $\mathcal{H}$ and $\mathcal{L}/M$ are respectively the convective operator – whose time scale is of order 1 – and the acoustic operator – whose time scale is of order M. To simplify the analysis, Schochet supposes that barotropic Euler system (8) is solved with periodic boundary conditions. Thus, Ω is the torus $\mathbb{T}^d := [a_1, b_1] \times \cdots \times [a_d, b_d]$ in $\mathbb{R}^d$ ($d \in \{1,2,3\}$). We now define

$$\begin{cases} \mathcal{E} = \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \in (L^2(\mathbb{T}^d))^{1+d} \text{ such that } \nabla r = 0 \text{ and } \nabla \cdot u = 0 \right\}, \\ \mathcal{E}^\perp = \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \in (L^2(\mathbb{T}^d))^{1+d} \text{ such that } \int_{\mathbb{T}^d} r dx = 0 \text{ and } \exists \phi \in H^1(\mathbb{T}^d), u = \nabla \phi \right\} \end{cases}$$

where the Hilbert space $(L^2(\mathbb{T}^d))^{1+d} := \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \text{ such that } \int_{\mathbb{T}^d} r^2 dx + \int_{\mathbb{T}^d} \|u\|^2 dx < +\infty \right\}$ is provided with the classical inner product $\langle q_1, q_2 \rangle = \int_{\mathbb{T}^d} q_1 q_2 dx$. We also recall the Hodge decomposition [23,24]:

Lemma 2.1.

$$\mathcal{E} \oplus \mathcal{E}^\perp = (L^2(\mathbb{T}^d))^{1+d} \quad \text{and} \quad \mathcal{E} \perp \mathcal{E}^\perp.$$

In other words, any $q \in (L^2(\mathbb{T}^d))^{1+d}$ can be decomposed into

$$q = \widehat{q} + q^\perp \quad \text{where} \quad (\hat{q}, q^\perp) \in \mathcal{E} \times \mathcal{E}^\perp \tag{10}$$

and this decomposition – named Hodge decomposition – is unique.

The Hodge decomposition (10) is also known as the Leray or Helmholtz decomposition. In the sequel, we define the Hodge projection $\mathbb{P}$ with $\mathbb{P}q := \widehat{q}$.

A simplified version of the result proposed by Schochet and applied to (8) (cf. theorems 2.1 and 3.1 in [19]) may be written in the following way:

Theorem 2.1. *Let $q(t,x)$ be solution of*

$$\begin{cases} \partial_t q + \mathcal{H}(q) + \dfrac{\mathcal{L}}{M}(q) = 0, \\ q(t=0,x) = q^0(x) \end{cases} \tag{11}$$

and let $\tilde{q}(t,x)$ be solution of

$$\begin{cases} \partial_t \tilde{q} + \mathbb{P}\mathcal{H}(\tilde{q}) = 0, \\ \tilde{q}(t=0,x) = \mathbb{P}q^0(x) \end{cases} \tag{12}$$

on $\Omega = \mathbb{T}^{d \in \{1,2,3\}}$. Then:

$$\begin{cases} \|q - \mathbb{P}q\|(t=0) = \mathcal{O}(M), \\ M \ll 1 \end{cases} \quad \Rightarrow \quad \|q - \tilde{q}\|(t \geqslant 0) = \mathcal{O}(M). \tag{13}$$

Let us note that estimate (13) is also written in [22] (see (2.127) in [22]). It is not difficult to verify that system (12) is equivalent to the incompressible Euler system with constant density

$$\begin{cases} \nabla \cdot u = 0, \\ \partial_t u + (u \cdot \nabla)u + \nabla \Pi = 0 \end{cases} \quad \text{and} \quad \begin{cases} r(t \geqslant 0, x) = C^{st}, \\ \nabla \cdot u(t = 0, x) = 0. \end{cases}$$

Thus, Theorem 2.1 means that when the initial conditions are well-prepared in the sense

$$\|q - \mathbb{P}q\|(t = 0) = \mathcal{O}(M), \tag{14}$$

the solution $q(t,x)$ of barotropic Euler system (11) is close to the solution $\tilde{q}$ of incompressible Euler system (12). Due to these statements, we will also consider $\mathcal{E}$ as the *well-prepared subspace*. Moreover, by using (7) and by supposing that the equation of state $p = p(\rho)$ is enough regular, we easily verify that condition (14) is equivalent to the condition

$$\begin{cases} \rho(t = 0, x) = \rho_* + \mathcal{O}(M^2)(x), \\ p(t = 0, x) = p_* + \mathcal{O}(M^2)(x), \\ u(t = 0, x) = \widehat{u}(x) + \mathcal{O}(M)(x) \end{cases} \quad \text{with} \quad \begin{cases} \rho_* = C_\rho^{st}, \\ p_* = p(C_\rho^{st}), \\ \nabla \cdot \widehat{u} = 0. \end{cases} \tag{15}$$

Condition (15) is identical to condition (5) except that there is a condition on the density in (15): this difference comes from the fact that (5) and (15) concern, respectively compressible Euler system (1) ($p = p(\rho, \varepsilon)$ in this case) and barotropic Euler system (6) ($p = p(\rho)$ in this case, which implies that the asymptotic expansions of ρ and p are necessarily linked).

2.3. Theoretical derivation in the linear case

The linear version of (11) is given by

$$\begin{cases} \partial_t q + Hq + \dfrac{L}{M} q = 0, \\ q(t = 0, x) = q^0(x) \end{cases} \tag{16}$$

where

$$\begin{cases} q = \begin{pmatrix} r \\ u \end{pmatrix}, \\ Hq = \begin{pmatrix} u_* \cdot \nabla r \\ (u_* \cdot \nabla)u \end{pmatrix} := (u_* \cdot \nabla)q, \\ Lq = a_* \begin{pmatrix} \nabla \cdot u \\ \nabla r \end{pmatrix}. \end{cases} \tag{17}$$

In (17), the velocity field u_* is a constant field in time and space. System (16) may also be deduced from a linearization of compressible Euler system (1) by defining r with

$$p(t,x) := \rho_* a_*^2 \left[1 + \frac{M}{a_*} r(t,x)\right] \tag{18}$$

instead of (7). Thus, in the sequel, r will be considered as a pressure perturbation rather than a density perturbation. Linear equation (16) was also studied in [8] in the Fourier space. Here, we study linear equation (16) in the physical space to obtain the linear version of Theorem 2.1. We use simple arguments based on the Hilbertian theory of linear operators. The proposed analysis seems to be more direct than the one proposed in [8] and, especially, allows to clearly identify the invariance property that will be the key argument to explain the inaccuracy of 2D or 3D Godunov type schemes at low Mach number in Section 4 and to propose a curative solution in Section 5. Let us now define the following energies

$$\begin{cases} E := \|q\|^2 = \text{total energy}, \\ E_{inc} := \|\hat{q}\|^2 = \text{incompressible energy}, \\ E_{ac} := \|q^\perp\|^2 = \text{acoustic (or compressible) energy}. \end{cases}$$

Of course, $E = E_{inc} + E_{ac}$ since $\mathcal{E} \perp \mathcal{E}^\perp$. The linear version of Theorem 2.1 is given by:

Proposition 2.1. *Let $q(t,x)$ be solution of*

$$\begin{cases} \partial_t q + Hq + \dfrac{L}{M} q = 0, \\ q(t = 0, x) = q^0(x), \end{cases} \tag{19}$$

let $\widetilde{q}(t,x)$ be solution of

$$\begin{cases} \partial_t \widetilde{q} + H\widetilde{q} = 0, \\ \widetilde{q}(t = 0, x) = \mathbb{P}q^0(x) \end{cases} \tag{20}$$

on $\Omega = \mathbb{T}^{d \in \{1,2,3\}}$ and let us define the Hodge decomposition $(\widehat{q}, q^\perp) \in \mathcal{E} \times \mathcal{E}^\perp$ of q. Then, $\widehat{q} = \widetilde{q}$ and $q^\perp$ is solution of (19) with the initial condition $q^{0\perp}$. Moreover, we have $E_{inc}(t \geqslant 0) = E_{inc}(t = 0)$ and $E_{ac}(t \geqslant 0) = E_{ac}(t = 0)$. This allows to write that

$$\|q-\mathbb{P}q\|(t=0)=\mathcal{O}(M) \quad \Rightarrow \quad \|q-\tilde{q}\|(t\geqslant 0)=\mathcal{O}(M). \tag{21}$$

Of course, Proposition 2.1 is a corollary of Theorem 2.1 (note that $H\tilde{q}=\mathbb{P}H\tilde{q}$ and that $\mathcal{H}(\tilde{q})\neq\mathbb{P}\mathcal{H}(\tilde{q})$). Nevertheless, since (19) is a constant-linear equation, we can propose a simple proof that will be useful to understand the inaccuracy of Godunov type schemes at low Mach number in Section 4 and to propose a curative solution in Section 5. This proof is based on two basic properties satisfied by linear equation (19). The first one is a classical energy conservation property:

Lemma 2.2. *Let $q(t,x)$ be solution of (19) on $\Omega=\mathbb{T}^{d\in\{1,2,3\}}$. Then:*

$$E(t\geqslant 0)=E(t=0).$$

The second one is an invariant property in the sense that $\mathcal{E}$ and $\mathcal{E}^{\perp}$ are invariant subspaces for Partial Differential Equation (PDE) (16). This property is central to understand the analysis proposed in Sections 4 and 5:

Lemma 2.3. *Let $q(t,x)$ be solution of (19) on $\Omega=\mathbb{T}^{d\in\{1,2,3\}}$. Then:*

(1) $\forall q^0\in\mathcal{E}:\ q(t\geqslant 0)\in\mathcal{E}$;
(2) $\forall q^0\in\mathcal{E}^{\perp}:\ q(t\geqslant 0)\in\mathcal{E}^{\perp}$.

Remark. Let us remark that (20) is equivalent to the system

$$\partial_t u+(u_*\cdot\nabla)u=0 \quad + \quad \begin{cases} r(t\geqslant 0,x)=C^{st}, \\ \nabla\cdot u(t=0,x)=0. \end{cases} \tag{22}$$

System (22) is "incompressible" since the invariance of $\mathcal{E}$ (see point 1 of Lemma 2.3) means that

$$\begin{cases} \partial_t u+(u_*\cdot\nabla)u=0, \\ \nabla\cdot u(t=0,x)=0 \end{cases} \quad \Rightarrow \quad \nabla\cdot u(t\geqslant 0,x)=0. \tag{23}$$

Let us note that (23) may be verified *via* a direct calculation on the PDE in (23).

Proof of Proposition 2.1: Let $q^0=\hat{q}^0+q^{0^{\perp}}$ be the Hodge decomposition of q^0 and let q_1 and q_2 be two solutions of (19) with the respective initial conditions $q_1(t=0,x)=\hat{q}^0(x)$ and $q_2(t=0,x)=q^{0^{\perp}}(x)$. By using Lemma 2.3, we obtain that $q_1(t\geqslant 0)\in\mathcal{E}$ and $q_2(t\geqslant 0)\in\mathcal{E}^{\perp}$. Moreover, by linearity of H and L, we find that $q=q_1+q_2$. Thus, by unicity of the Hodge decomposition (10), we conclude that $\hat{q}=q_1$ and $q^{\perp}=q_2$. And since $\mathcal{E}=KerL$, we find that $\tilde{q}=\hat{q}$. We conclude the proof by using Lemma 2.2. □

It remains to prove Lemmas 2.2 and 2.3:

Proof of Lemma 2.2: We have $\frac{d}{dt}||q||^2=2\langle q,\partial_t q\rangle=-2\langle q,Hq\rangle-2\langle q,\frac{L}{M}q\rangle$. We have also

$$\begin{aligned} 2\langle q,Hq\rangle &= \int_{\mathbb{T}^d}u_*\cdot\nabla\|q\|^2dx=\int_{\mathbb{T}^d}\nabla\cdot(\|q\|^2u_*)dx-\int_{\mathbb{T}^d}\|q\|^2\nabla\cdot u_*dx=\int_{\partial\mathbb{T}^d}\|q\|^2u_*\cdot nds-\int_{\mathbb{T}^d}\|q\|^2\nabla\cdot u_*dx \\ &= 0-\int_{\mathbb{T}^d}\|q\|^2\nabla\cdot u_*dx=0 \quad (\nabla\cdot u_*=0 \text{ since } u_*=C^{st}). \end{aligned}$$

Moreover $\langle q,Lq\rangle=a_*(\int_{\mathbb{T}^d}r\nabla\cdot udx+\int_{\mathbb{T}^d}u\cdot\nabla rdx)=a_*\int_{\partial\mathbb{T}^d}ru\cdot nds=0$. □

The proof of Lemma 2.3 is based on the following lemma whose proof is in Appendix A:

Lemma 2.4. *Let $\mathcal{A}$ be a linear isometry on a Hilbert space $\mathbb{H}$ and let $\mathcal{E}$ be a vector subspace of $\mathbb{H}$. Then:*

$$\mathcal{A}\mathcal{E}=\mathcal{E} \quad \Rightarrow \quad \mathcal{A}\mathcal{E}^{\perp}\subset\mathcal{E}^{\perp}.$$

Proof of Lemma 2.3: Let us define the linear application

$$\mathcal{A}_t^{(H,L/M)}: \quad \begin{cases} (L^2(\mathbb{T}^d))^{1+d}\to(L^2(\mathbb{T}^d))^{1+d}, \\ q^0\mapsto\mathcal{A}_t^{(H,L/M)}q^0:=q(t,\cdot) \end{cases} \tag{24}$$

where q is solution of (19) on $\Omega=\mathbb{T}^d$. Lemma 2.2 allows us to write that $\mathcal{A}_t^{(H,L/M)}$ is an isometry. Thus, by using Lemma 2.4, we just have to prove that $\mathcal{A}_t^{(H,L/M)}$ is surjective on $\mathcal{E}$ to prove that $\mathcal{E}^{\perp}$ is invariant. Since u_* is a constant velocity field in time and space, we have $\mathcal{A}_t^{(H,L/M)}=\mathcal{T}_t^H\circ\mathcal{A}_t^{(0,L/M)}$ where $\mathcal{T}_t^H$ is the application associated to the linear transformation $q(t,x)\mapsto q(t,x-u_*t)$. Moreover, $\mathcal{A}_t^{(0,L/M)}$ is the identity application on $\mathcal{E}$ since $\mathcal{E}=KerL$. This means that $\mathcal{A}_t^{(H,L/M)}=\mathcal{T}_t^H$ on $\mathcal{E}$. Thus, we obtain that $\mathcal{A}_t^{(H,L/M)}$ is surjective on $\mathcal{E}$ by noting that $\mathcal{T}_t^H$ is surjective on $\mathcal{E}$. Let us remark that $\mathcal{A}_t^{(H,L/M)}\neq\mathcal{T}_t^H\circ\mathcal{A}_t^{(0,L/M)}$ when we have only $\nabla\cdot u_*(x)=0$. That is why we impose $u_*=C^{st}$ in (17). □

2.4. A sufficient condition to avoid the creation of spurious waves at low Mach number

Estimate (21) – that characterizes the convergence of a compressible flow toward an incompressible flow in the linear case – is a direct consequence of Lemma 2.1 (energy conservation property) and of Lemma 2.3 (invariance property of the well-prepared subspace $\mathcal{E}$ and of the acoustic subspace $\mathcal{E}^{\perp}$). Since $\tilde{q} = \mathbb{P}q$ in the case of linear equation (19), estimate (21) may also be written with

$$\|q - \mathbb{P}q\|(t = 0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M). \tag{25}$$

If we just want to prove estimate (25) without using the property $\tilde{q} = \mathbb{P}q$, only the well-posedness of linear equation (19) and the invariance of $\mathcal{E}$ are important. Indeed, we have the following result:

Theorem 2.2. *Let $q(t,x)$ be solution of the linear PDE*

$$\begin{cases} \partial_t q + \mathcal{F}_x q = 0, \\ q(t = 0) = q^0 \end{cases} \tag{26}$$

supposed to be well-posed on $L^{\infty}([0,+\infty[;(L^2(\mathbb{T}^d))^{1+d})$ where $\mathcal{F}_x$ is a linear spatial differential operator. Then:

(1) the solution $q(t,x)$ satisfies

$$\|q - \mathbb{P}q\|(t = 0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \tilde{q}\|(t \geqslant 0) = \mathcal{O}(M) \tag{27}$$

where $\tilde{q}(t,x)$ is solution of (26) with the initial condition $\tilde{q}(t = 0) = \mathbb{P}q^0$. Nevertheless, we do not have a priori $\|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M)$;

(2) when $\mathcal{F}_x$ is such that

$$\forall q^0 \in \mathcal{E} : \ q(t \geqslant 0) \in \mathcal{E}, \tag{28}$$

the solution $q(t,x)$ satisfies (27) and

$$\|q - \mathbb{P}q\|(t = 0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M). \tag{29}$$

Thus, we have also $\|\tilde{q} - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M)$.

The second point of Theorem 2.2 is important since estimate (29) means that $q(t \geqslant 0)$ is almost in the well-prepared subspace $\mathcal{E}$ when it is the case at $t = 0$. In other words, under sufficient condition (28) and when the initial conditions are well-prepared, PDE (26) does not create any spurious wave. At the opposite, the first point of Theorem 2.2 is not really important since estimate (27) does not mean that $q(t \geqslant 0)$ is almost in the well-prepared subspace $\mathcal{E}$. This first point is only mentioned to underline the importance of the invariance of $\mathcal{E}$. Moreover, we underline the fact that we do not impose that the energy $E(t)$ is a constant or a decreasing function (nevertheless, $E(t \geqslant 0)$ is bounded since (26) is supposed to be well-posed). Let us also underline that $\tilde{q} \neq \mathbb{P}q$ *a priori*. We would have $\tilde{q} = \mathbb{P}q$ by imposing also the invariance of $\mathcal{E}^{\perp}$. At last, when $\mathcal{F}_x = H + \frac{L}{M}$, (29) implies (21) since we have in that case $\tilde{q} = \mathbb{P}q$ ($\mathcal{E}^{\perp}$ is invariant) and $\mathcal{F}_x|_{\mathcal{E}} = H$ ($\mathcal{E} = KerL$). Of course, when $\mathcal{F}_x = \frac{L}{M}$, we have $\tilde{q} = \mathbb{P}q = \mathbb{P}q^0$ and (29) implies

$$\|q - \mathbb{P}q\|(t = 0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q^0\|(t \geqslant 0) = \mathcal{O}(M). \tag{30}$$

Let us now define a non-linear X scheme of Godunov type (X = Roe [1] for example) or not (X = kinetic scheme [20] for example) applied to compressible Euler system (1) or to barotropic Euler system (6) on any mesh. Theorem 2.2 gives us a simple and clear theoretical framework:

- **firstly to understand** (at least partly) on a Cartesian mesh the origin of the inaccuracy of the non-linear X scheme at low Mach number by analyzing the first-order modified equation

$$\begin{cases} \partial_t q + \mathcal{F}_x^{X,L/M} q = 0, \\ q(t = 0) = q^0 \end{cases} \tag{31}$$

associated to the X scheme applied to the linear wave equation

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = 0, \\ q(t = 0) = q^0. \end{cases} \tag{32}$$

Indeed, estimate (29) is a necessary condition to be accurate at low Mach number. And, if sufficient condition (28) is not satisfied for (31), estimate (29) may not be satisfied, which means that it may exist spurious waves characterized by the estimate $\|q - \mathbb{P}q\|(\tau_{ac}) \gg \mathcal{O}(M)$, τ_{ac} being an acoustic time scale of order $\mathcal{O}(M)$. For example, we show in Section 4 that condition (28) is not satisfied in 2D and 3D when $\mathcal{F}_x := \mathcal{F}_x^{Godunov,L/M}$ (see point 2 of Lemma 4.2) and that there exists 2D and 3D initial

conditions which create spurious waves (see point 2 of Proposition 4.1). As a consequence, any non-linear Godunov scheme cannot be accurate at low Mach number. Let us underline that condition (28) concerns Eq. (31) that is linear. This means that the origin of the inaccuracy of non-linear Godunov type schemes at low Mach number can be explained (at least partly) on a Cartesian mesh with linear arguments and without any convection term (indeed, $H = 0$ in $\mathcal{F}_x^{Godunov,L/M}$);

- **secondly to propose** a new non-linear X scheme – named *low Mach X scheme* in Section 5 (see Conjecture 5.1) – in such a way condition (28) is satisfied for the new first-order modified equation

$$\begin{cases} \partial_t q + \mathcal{F}_x^{low\ Mach\ X,L/M} q = 0, \\ q(t=0) = q^0 \end{cases} \tag{33}$$

associated to this *low Mach X scheme* applied to (32), although (28) is not satisfied for (31). When X is a Godunov type scheme, we propose in Section 5 a simple modification of the X scheme to obtain such an operator $\mathcal{F}_x^{low\ Mach\ X,L/M}$. We emphasize that this modification can be applied to any colocated X scheme and on any mesh (see Conjecture 5.1). Numerical results proposed in Section 7 show that these *low Mach X schemes* are stable and accurate at low Mach number when X = Roe [1] and X = VFRoe [2,3] (see also Section 5.2 for a stability analysis of the linear *low Mach Godunov scheme*). Moreover, we show in Section 5.5 that other colocated schemes of Godunov type [6–10,16–18] or not [12] that are accurate at low Mach number are similar or identical to a *low Mach X scheme* when the Mach number goes to zero.

Proof of Theorem 2.2: Estimate (27) is a direct consequence of the well-posedness of PDE (26). Let us now define $\bar{q}(t,x)$ solution of (26) with the initial condition $\bar{q}(t=0,x) := q^0 - \mathbb{P}q^0$. By linearity, we have $q = \tilde{q} + \bar{q}$. Moreover

$$\|q - \mathbb{P}q\| = \|\tilde{q} - \mathbb{P}\tilde{q} + \bar{q} - \mathbb{P}\bar{q}\| = \|\bar{q} - \mathbb{P}\bar{q}\|$$

by using the invariance property (28). Then, we have

$$\|q - \mathbb{P}q\| \leqslant \|\bar{q}\|$$

since $(\mathbf{1} - \mathbb{P})$ is an orthogonal projection. We conclude the proof by using again the well-posedness of (26) which implies that $\|\bar{q}\| = \mathcal{O}(M)$ when $\|\bar{q}^0\| := \|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M)$. □

2.5. Two additional conditions to obtain a good colocated low Mach scheme

Estimate (29) means that PDE (33) does not create any spurious wave. Nevertheless, it remains to verify *a posteriori* that:

- **firstly**, the new X scheme – whose first-order modified equation is given by (33) – is obtained from the original X scheme *in a simple way* on any mesh;
- **secondly**, the energy in the well-prepared subspace $\mathcal{E}$ (i.e. the incompressible energy) *is not diffused with a time scale of order M*. Indeed, to be accurate at low Mach number, we need to control the energy in $\mathcal{E}^{\perp}$ to avoid the creation of spurious waves through estimate (29) but we need *also* to control the dissipation of the incompressible energy in the well-prepared subspace $\mathcal{E}$.

The *low Mach X schemes* proposed in Conjecture 5.1 satisfy these two additional conditions. We will propose in Section 5.6 two other schemes whose first-order modified equations satisfy sufficient condition (28) but that do not satisfy the first or the second additional condition. In Section 6, we will also propose a scheme whose first-order modified equation satisfies a sufficient condition that is weaker than (28) but that does not satisfy the first additional condition.

3. Formal analysis of the Roe scheme at low Mach number on a Cartesian mesh

Some authors tried to give a formal analysis of the inaccuracy of Godunov type schemes at low Mach number [4,6–8,10]. All these formal analysis use an asymptotic expansion based on the Mach number. Although the formal analysis in [4,6,10] allows to find the origin of the inaccuracy, its arguments seem to be inexact since they concern the 1D-case that, in fact, does not suffer of any inaccuracy at low Mach number as it is underlined in [5] (see p. 362 in [5]). We will justify this assertion with formal arguments in this section and with theoretical arguments in Section 4. In [8], the formal analysis is more accurate. Nevertheless, the importance of the space dimension is not clearly underlined since the inaccuracy of the Godunov type scheme is explained with an analysis of a 1D-Riemann problem applied to a perfect gas at low Mach number. In [7], the inaccuracy is studied in the case of the Roe scheme [1] with 2D arguments (see lemma 3.1 and proposition 3.1 p. 72-73 in [7]). Here, we will propose a formal analysis in the case of the 2D Roe scheme when the mesh is Cartesian. This analysis is more accurate than the one proposed in [4,6,10] and is more simple than the one proposed in [7,8]. Let us note that the particular case of triangular (or tetrahedral) mesh is mentioned in Section 5.8. Thus, in this section, we firstly recall the formal approach proposed in [6,10] and we clearly underline its defaults. Secondly, we propose a formal approach that clearly takes into account the space dimension and we also discuss about the 1D-formal analysis proposed in [4,8]. This formal approach may be seen as an introduction to the theoretical approach proposed in Section 4.

In the sequel, the mesh is a 2D Cartesian mesh. The subscript (i,j) defines the center of each cell of the Cartesian mesh, $(i\pm 1/2,j)$ and $(i,j\pm 1/2)$ defining the interfaces of the cell (i,j). Moreover, Δx_k is the space step in the direction x_k ($k\in\{1,2\}$). The formal arguments proposed below would be identical for a 3D Cartesian mesh.

3.1. A first formal approach

We show in this subsection why the argument proposed in [6,10] to justify the inaccuracy of Godunov type schemes is not correct although it allows to find the origin of the inaccuracy. The Roe scheme [1] applied to compressible Euler system (1) (written with physical dimensions) on a 2D Cartesian mesh is given by

$$\begin{cases} \frac{d}{dt}U_{i,j}+\dfrac{F_{1,i+1/2,j}-F_{1,i-1/2,j}}{\Delta x_1}+\dfrac{F_{2,ij+1/2}-F_{2,ij-1/2}}{\Delta x_2}=0,\\ U_{i,j}(t=0)=U_{ij}^0 \end{cases} \tag{34}$$

where

$$\begin{cases} F_{1,i+1/2,j}=\dfrac{\mathcal{F}_1(U_{i,j})+\mathcal{F}_1(U_{i+1,j})}{2}-\dfrac{1}{2}|A_1(U_{i+1/2,j})|(U_{i+1,j}-U_{i,j}), & \text{(a)}\\ F_{2,ij+1/2}=\dfrac{\mathcal{F}_2(U_{i,j})+\mathcal{F}_2(U_{ij+1})}{2}-\dfrac{1}{2}|A_2(U_{ij+1/2})|(U_{ij+1}-U_{i,j}). & \text{(b)} \end{cases} \tag{35}$$

In (34) and (35), $U:=(\rho,\rho u,\rho E)^T$ with $u=(u_k)_{k=1,\dots,d}\in\mathbb{R}^d$, $\mathcal{F}_k(U)$ is the flux of (1) in the direction k and $A_k(U)=\frac{D\mathcal{F}_k}{DU}(U)(k\in\{1,2\})$. The quantities $U_{i+1/2,j}$ and $U_{ij+1/2}$ are defined with the Roe average. This average is given in [1] for a perfect gas and in [25] for any equation of state $p(\rho,\varepsilon)$ (which admits an entropy). Nevertheless, as there are no shock waves at low Mach number, we do not need to satisfy any property related, for example, to the preservation of a stationary shock wave (this is a property that justifies the choice of the Roe average for compressible flows). Thus, for the sake of simplicity, we can suppose for example that $U_{i+1/2,j}:=(U_{i,j}+U_{i+1,j})/2$ and $U_{ij+1/2}:=(U_{i,j}+U_{ij+1})/2$. This choice does not change formulae (35) and formulae (37) and (38) below. The important point is just to note that we have $\mathcal{O}(\|U_{i+1/2,j}\|)=\mathcal{O}(\|U_{i,j}\|)=\mathcal{O}(\|U_{i+1,j}\|)$ ($\|\cdot\|$ is the l^2-norm) for any choice of the average because of the regularity of the solution at low Mach number. This means that we can define a global Mach number M such that

$$\forall(i,j):\ \mathcal{O}\left(\frac{\|u_{i+1/2,j}\|}{a_{i+1/2,j}}\right)=\mathcal{O}\left(\frac{\|u_{ij+1/2}\|}{a_{ij+1/2}}\right)=\mathcal{O}\left(\frac{\|u_{ij}\|}{a_{ij}}\right)=M\ll 1. \tag{36}$$

We will use this property below to analyse the different orders of magnitude of the Roe scheme. When the flow is subsonic (i.e. $\|u_{i+1/2,j}\|\leqslant a_{i+1/2,j}$), formula (35)(a) may be written with [5]

$$\begin{cases} u_{1,i+1/2,j}>0: & F_{1,i+1/2,j}=\mathcal{F}_1(U_i)+(u_1-a)_{i+1/2,j}\cdot(\Delta U)^-_{i+1/2,j},\\ u_{1,i+1/2,j}\leqslant 0: & F_{1,i+1/2,j}=\mathcal{F}_1(U_{i+1})-(u_1+a)_{i+1/2,j}\cdot(\Delta U)^+_{i+1/2,j} \end{cases} \tag{37}$$

knowing that

$$(\Delta U)^{\pm}_{i+1/2,j}:=\left(\frac{[p]\pm\rho a[u_1]}{2a^2}\right)_{i+1/2,j}\cdot\begin{pmatrix}1\\ u_1\pm a\\ u_2\\ \mathbf{H}\pm u_1 a\end{pmatrix}_{i+1/2,j} \tag{38}$$

with $[f]_{i+1/2,j}:=f_{i+1,j}-f_{i,j}$. In (38), $\mathbf{H}:=E+p/\rho$ is the total enthalpy. Of course, we have symmetric formulae for $F_{2,ij+1/2}$. Let us now define $p_{i+1/2}:=(\rho u^2+p)_{i+1/2}-(\rho u)_{i+1/2}\cdot u_{i+1/2}$. We deduce from (37) that for any 1D subsonic flow

$$\begin{cases} u_{i+1/2}>0: p_{i+1/2}=\dfrac{p_i+p_{i+1}}{2}-\left(\dfrac{\rho a[u]}{2}\right)_{i+1/2}+\rho_i u_i(u_i-u_{i+1/2})-\left(\dfrac{|u|}{2a}([p]-\rho a[u])\right)_{i+1/2},\\ u_{i+1/2}\leqslant 0: p_{i+1/2}=\dfrac{p_i+p_{i+1}}{2}-\left(\dfrac{\rho a[u]}{2}\right)_{i+1/2}+\rho_{i+1} u_{i+1}(u_{i+1}-u_{i+1/2})-\left(\dfrac{|u|}{2a}([p]+\rho a[u])\right)_{i+1/2}. \end{cases} \tag{39}$$

Thus, by noting $M_{i+1/2}:=\left(\dfrac{|u|}{a}\right)_{i+1/2}$ and by replacing $\left(\dfrac{\rho a[u]}{2}\right)_{i+1/2}$ with $\left(\dfrac{\rho|u|[u]}{2M}\right)_{i+1/2}$, we obtain

$$p_{i+1/2}=\frac{p_i+p_{i+1}}{2}-\left(\frac{\rho|u|[u]}{2M}\right)_{i+1/2}+\cdots. \tag{40}$$

By using (36), we can say that $M_{i+1/2}\ll 1$. In [10], the inaccuracy of the Roe scheme is explained by noting that

$$\left(\frac{\rho|u|[u]}{2M}\right)_{i+1/2}\to+\infty\quad\text{when}\quad M_{i+1/2}\to 0^+. \tag{41}$$

Indeed, (41) seems to show that the pressure $p_{i+1/2}$ deviates dramatically from the arithmetic average of p_i and p_{i+1} when $M_{i+1/2} \to 0^+$. In [6], the proposed argument is equivalent to (41) (see (12) in [6]). Thus, it is proposed in [10] to correct the Roe scheme by replacing fluxes (35) in (34) with the corrected fluxes

$$\begin{cases} F^{corrected}_{1,i+1/2,j} = F_{1,i+1/2,j} + \dfrac{(\rho a)_{i+1/2,j}}{2} \begin{pmatrix} 0 \\ [u_1] \\ 0 \\ 0 \end{pmatrix}_{i+1/2,j}, \\ F^{corrected}_{2,i,j+1/2} = F_{2,i,j+1/2} + \dfrac{(\rho a)_{i,j+1/2}}{2} \begin{pmatrix} 0 \\ 0 \\ [u_2] \\ 0 \end{pmatrix}_{i,j+1/2}. \end{cases} \tag{42}$$

The Roe scheme with pressure correction (42) gives good numerical results. Let us note that the pressure correction could be applied to any Godunov type scheme. Nevertheless, (41) is not correct and, thus, does not justify the correction (42): indeed, (41) does not take into account the fact that $|\rho a u| \ll p$ at low Mach number since $\mathcal{O}(p) = \mathcal{O}(\rho a^2)$. Thus, (41) should be replaced by

$$\left(\frac{\rho |u| [u]}{2M}\right)_{i+1/2} \ll \frac{p_i + p_{i+1}}{2} \quad \text{when} \quad M_{i+1/2} \to 0^+. \tag{43}$$

Moreover, the previous formal analysis (partly inspired from [6,10]) does not take into account the space dimension since it is an 1D analysis (1D arguments are also used in [4,8,26]): we will see in the sequel that the space dimension is essential to analyse Godunov type schemes at low Mach number from a formal or theoretical point of view.

3.2. A more precise formal approach: importance of the space dimension

We propose in this subsection another formal approach that is more precise than the one proposed in [6,10]. This formal approach is similar to the one proposed in [4,8]. Nevertheless, it is more precise since we clearly underline that the space dimension is essential to formally analyse the Roe scheme, which is not the case in [4,8] (as in [6,10], 1D arguments are used in [4,8]). Moreover, the proposed formal analysis is more simple and general than the one proposed in [8]. Indeed, firstly, the formal approach in [8] is based on an explicit resolution of a 1D-Riemann problem and, secondly, it supposes that the fluid is a perfect gas. Here, we do not solve such 1D-Riemann problem and the arguments are correct for any admissible equation of state. The dimensionless version of (39) is given by

$$\begin{cases} u_{i+1/2} > 0 : p_{i+1/2} = \dfrac{p_i + p_{i+1}}{2} - M\left(\dfrac{\rho a [u]}{2}\right)_{i+1/2} + M^2 \rho_i u_i (u_i - u_{i+1/2}) - M\left(\dfrac{|u|}{2a}([p] - M\rho a[u])\right)_{i+1/2} \\ u_{i+1/2} \leqslant 0 : p_{i+1/2} = \dfrac{p_i + p_{i+1}}{2} - M\left(\dfrac{\rho a [u]}{2}\right)_{i+1/2} + M^2 \rho_{i+1} u_{i+1} (u_{i+1} - u_{i+1/2}) - M\left(\dfrac{|u|}{2a}([p] + M\rho a[u])\right)_{i+1/2} \end{cases} \tag{44}$$

with now $p_{i+1/2} := (M^2 \rho u^2 + p)_{i+1/2} - M^2 (\rho u)_{i+1/2} \cdot u_{i+1/2}$. Thus, at low Mach number, we have

$$p_{i+1/2} = \frac{p_i + p_{i+1}}{2} - M\left(\frac{\rho a [u]}{2}\right)_{i+1/2} - M\left(\frac{|u|}{2a}[p]\right)_{i+1/2} + \mathcal{O}(M^2 \Delta x). \tag{45}$$

Let us note that (45) is compatible with (43) and not with (41) (we emphasize that u in (45) is dimensionless which is not the case in (41) and (43); on the other hand, the Mach number M in (45) is of the same order as the Mach number $M_{i+1/2}$ in (41) and (43) because of (36)). Let us suppose that the initial conditions are well-prepared in the sense of (5). Thus, we have $M\left(\frac{|u|}{2a}[p]\right)_{i+1/2} = \mathcal{O}(M^3 \Delta x)$. Moreover, in 1D, $\nabla \cdot \hat{u} = 0$ is equivalent to $\widehat{u}(x) = C^{st}$: in that case, $[u]_{i+1/2} = \mathcal{O}(M \Delta x)$. This means that

$$M\left(\frac{\rho a [u]}{2}\right)_{i+1/2} = \mathcal{O}(M^2 \Delta x). \tag{46}$$

Thus, (45) coupled to (5) means that

$$p_{i+1/2}(t = 0) = p_* + \mathcal{O}(M^2 \Delta x) + \mathcal{O}(M^2). \tag{47}$$

Then, because of the pressure gradient $\nabla p / M^2$ in momentum equation (1)(b), the 1D velocity field should be such that

$$\frac{d}{dt} u_i(t) = \mathcal{O}(\Delta x) + \mathcal{O}(1) \tag{48}$$

where $\mathcal{O}(\Delta x) + \mathcal{O}(1)$ is a (discrete) non-constant field. Then, we have

$$u_i(\Delta t) = \widehat{u}_i + \mathcal{O}(M) + \Delta t[\mathcal{O}(\Delta x) + \mathcal{O}(1)] + \mathcal{O}(\Delta t^2)$$

when $\Delta t \ll 1$. And when $\Delta t = \tau_{ac}$ with $\tau_{ac} = \mathcal{O}(M)$ (τ_{ac} is an acoustic time scale), we may formally write that

$$u_i(\tau_{ac}) = \hat{u}_i + \mathcal{O}(M\Delta x) + \mathcal{O}(M) \tag{49}$$

where $\mathcal{O}(M\Delta x) + \mathcal{O}(M)$ is a (discrete) non-constant velocity field. In that case, the 1D pressure field should be such that

$$p_i(\tau_{ac}) = p_* + \mathcal{O}(M^2\Delta x) + \mathcal{O}(M^2) \tag{50}$$

by using (2)(c). In [4], it is abruptly written that in1D, the pressure field satisfies $p_i(t > 0) = p* + \mathcal{O}(M)$ (see Eq. (30) in [4]): this estimate is in contradiction with (50). In [8], the analysis of the 1D-Riemann problem for a perfect gas leads to an equation similar to (45) (see Eq. (43) in [8]). Nevertheless, as in [4], it is deduced from the 1D analysis that "*Godunov solver contains, after one time step, acoustic pressure waves of order M that are much larger than the pressure fluctuations due to the incompressible component*" (cf. p. 666 in [8]). Numerical results are also proposed in 1D to justify this assertion (cf. Fig. 4 p. 667 in [8]). These numerical results are correct. Nevertheless, they cannot justify the existence of the inaccuracy of Godunov type schemes at low Mach number. Indeed, the initial condition used to obtain the 1D numerical results in [8] is not well-prepared since the initial velocity field has a fluctuation of order 1 and not of order M. As a consequence, in (45), $M\left(\frac{\rho a[u]}{2}\right)_{i+1/2} = \mathcal{O}(M\Delta x)$ instead of (46). In that case, estimate (50) induced by any 1D well-prepared initial condition should be replaced by the estimate $p_i(\tau_{ac}) = p* + \mathcal{O}(M\Delta x)$ (see below the 2D case for the details). In fact, the formal analysis in [4,8] cannot be satisfactory because, as in [6,10], they use 1D arguments (an 1D argument is also used in [26]: see estimate (11) in [26]). As a consequence, any divergence-free field is trivial and it is impossible to clearly understand the origin of the inaccuracy of Godunov type schemes at low Mach number. Thus, even for a formal analysis, we have to take into account the dimension of the space to analyse the inaccuracy of Godunov type schemes at low Mach number.

The 2D version of (45) is given by

$$\begin{cases} p_{i+1/2,j} = \dfrac{p_{ij} + p_{i+1,j}}{2} - M\left(\dfrac{\rho a[u_1]}{2}\right)_{i+1/2,j} - M\left(\dfrac{|u_1|}{2a}[p]\right)_{i+1/2,j} + \mathcal{O}(M^2\Delta x), \\ p_{ij+1/2} = \dfrac{p_{ij} + p_{ij+1}}{2} - M\left(\dfrac{\rho a[u_2]}{2}\right)_{ij+1/2} - M\left(\dfrac{|u_2|}{2a}[p]\right)_{ij+1/2} + \mathcal{O}(M^2\Delta x) \end{cases} \tag{51}$$

(for the sake of simplicity, we choose $\Delta x_1 = \Delta x_2 := \Delta x$) with

$$\begin{cases} p_{i+1/2,j} := (M^2\rho u_1^2 + p)_{i+1/2,j} - M^2(\rho u_1)_{i+1/2,j} \cdot u_{1,i+1/2,j}, \\ p_{ij+1/2} := (M^2\rho u_2^2 + p)_{ij+1/2} - M^2(\rho u_2)_{ij+1/2} \cdot u_{2,ij+1/2}. \end{cases}$$

In 2D, $\nabla \cdot \widehat{u} = 0$ is not equivalent to $\hat{u}(x) = C^{st}$. Thus, when the initial conditions are well-prepared in the sense of (5), we have *a priori*

$$\begin{cases} M\left(\dfrac{\rho a[u_1]}{2}\right)_{i+1/2,j} = \mathcal{O}(M\Delta x), \\ M\left(\dfrac{\rho a[u_2]}{2}\right)_{ij+1/2} = \mathcal{O}(M\Delta x). \end{cases} \tag{52}$$

We deduce from (51) and (52) that the 2D version of (47) should be such that

$$\begin{cases} p_{i+1/2,j}(t=0) = p_* + \mathcal{O}(M\Delta x) + \mathcal{O}(M^2), & \text{(a)} \\ p_{ij+1/2}(t=0) = p_* + \mathcal{O}(M\Delta x) + \mathcal{O}(M^2) & \text{(b)}. \end{cases} \tag{53}$$

Thus, the 2D version of the 1D estimate (48) is given by

$$\frac{d}{dt}u_{ij}(t) = \mathcal{O}\left(\frac{\Delta x}{M}\right) + \mathcal{O}(1)$$

where $\mathcal{O}(\frac{\Delta x}{M}) + \mathcal{O}(1)$ is a (discrete) non-divergence-free field. As a consequence, the 2D version of the 1D estimate (49) is now given by

$$u_{ij}(\tau_{ac}) = \hat{u}_{ij} + \mathcal{O}(\Delta x) + \mathcal{O}(M) \tag{54}$$

(with again $\tau_{ac} = \mathcal{O}(M)$) where $\mathcal{O}(\Delta x) + \mathcal{O}(M)$ is a non-divergence-free field, although $u_{ij}(t=0) = \hat{u}_{ij} + \mathcal{O}(M)$. In that case, the 2D pressure field should be

$$p_{ij}(\tau_{ac}) = p_* + \mathcal{O}(M\Delta x) + \mathcal{O}(M^2) \tag{55}$$

although $p_{ij}(t=0) = p_* + \mathcal{O}(M^2)$. The velocity field $\mathcal{O}(\Delta x)$ in (54) and the pressure perturbation $\mathcal{O}(M\Delta x)$ in (55) are spurious waves that make impossible the convergence when $M \to 0$ (with a given Δx) of a 2D (or 3D) discrete compressible solution toward a 2D (or 3D) discrete incompressible solution when compressible Euler system (1) is discretized with a Roe scheme.

Indeed, Theorem 2.1 and Proposition 2.1 underline that a compressible flow converges toward an incompressible flow at low Mach number if and only if the initial conditions are well-prepared in the sense of (5). More precisely, any $\mathcal{O}(1)$ (or $\mathcal{O}(\Delta x)$ at the discrete level) velocity fluctuation that is a non-divergence-free field induces a $\mathcal{O}(M)$ (or $\mathcal{O}(M\Delta x)$) pressure fluctuation, and *vice versa*. On the other hand, estimates (54) and (55) seem to say that when the mesh is such that $\Delta x = \mathcal{O}(M)$, there are no spurious waves: of course, the condition $\Delta x = \mathcal{O}(M)$ is impracticable in 2D (or 3D) geometry when $M \to 0$. In a way, we can say that the *compressible* Godunov scheme converges to an *incompressible* colocated scheme under the consistency condition $\Delta x = \mathcal{O}(M)$. At the opposite, estimates (49) and (50) show that the 1D Roe scheme should not create any spurious waves as soon as the initial conditions are well-prepared in the sense of (5): as a consequence, the 1D Roe scheme should not suffer of any particular inaccuracy when the Mach number goes to zero.

Finally, let us note that estimates (49) and (50) mean that

$$\|q - \mathbb{P}q\|(\tau_{ac}) = \mathcal{O}(M) \tag{56}$$

at the discrete level where $q := (r,u)^T$ with r defined by (18). In the same way, estimates (54) and (55) mean that

$$\|q - \mathbb{P}q\|(\tau_{ac}) \gg \mathcal{O}(M) \tag{57}$$

at the discrete level when $\Delta x \gg \mathcal{O}(M)$. In Section 4, we will clearly justify in the linear case from a theoretical point of view estimates (56) and (57): see estimates (65) and (66) in Proposition 4.1.

4. Analysis with the linear wave equation

We now analyse the inaccuracy of Godunov type schemes at low Mach number with the approach proposed in Section 2.4 by studying the linear wave equation

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = 0, \\ q(t=0,x) = q^0(x) \end{cases} \tag{58}$$

solved on $\Omega = \mathbb{T}^{d\in\{1,2,3\}}$ with the Godunov scheme (in the linear case, all Godunov type schemes are identical). We recall that (58) satisfies the estimate

$$\|q - \mathbb{P}q\|(t=0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M) \tag{59}$$

(see point 2 of Theorem 2.2). This linear approach will justify the formal approach proposed in Section 3.2.

4.1. The Godunov scheme and its first-order modified equation

The Godunov scheme applied to linear wave equation (58) on a 2D Cartesian mesh is given by

$$\begin{cases} \frac{d}{dt} r_{i,j} + \dfrac{a_*}{M} \cdot \left(\dfrac{u_{1,i+1,j} - u_{1,i-1,j}}{2\Delta x_1} + \dfrac{u_{2,i,j+1} - u_{2,i,j-1}}{2\Delta x_2} \right) = \dfrac{a_*}{2M} \cdot \left(\dfrac{r_{i+1,j} - 2r_{i,j} + r_{i-1,j}}{\Delta x_1} + \dfrac{r_{i,j+1} - 2r_{i,j} + r_{i,j-1}}{\Delta x_2} \right), \\ \frac{d}{dt} u_{1,i,j} + \dfrac{a_*}{M} \cdot \dfrac{r_{i+1,j} - r_{i-1,j}}{2\Delta x_1} = \dfrac{a_*}{2M} \cdot \dfrac{u_{1,i+1,j} - 2u_{1,i,j} + u_{1,i-1,j}}{\Delta x_1}, \\ \frac{d}{dt} u_{2,i,j} + \dfrac{a_*}{M} \cdot \dfrac{r_{i,j+1} - r_{i,j-1}}{2\Delta x_2} = \dfrac{a_*}{2M} \cdot \dfrac{u_{2,i,j+1} - 2u_{2,i,j} + u_{2,i,j-1}}{\Delta x_2}. \end{cases} \tag{60}$$

We recall that $q := \begin{pmatrix} r \\ u \end{pmatrix}$ where r is the pressure perturbation defined with (18). For the sake of simplicity, we suppose in the sequel that $\forall k : \Delta x_k = \Delta x$. Thus, the first-order modified equation associated to (60) is given by

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = B_\kappa q, \\ q(t=0,x) = q^0(x) \end{cases} \tag{61}$$

with

$$B_\kappa q = \mathcal{K} \begin{pmatrix} \Delta r \\ \frac{\partial^2 u_1}{\partial x_1^2} \\ \frac{\partial^2 u_2}{\partial x_2^2} \end{pmatrix} \quad \text{and} \quad \mathcal{K} = \begin{pmatrix} \nu_r & 0 & 0 \\ 0 & \nu_{u_1} & 0 \\ 0 & 0 & \nu_{u_2} \end{pmatrix} \tag{62}$$

where $\nu_r = \nu_{u_k} = \nu_{num} := a_* \frac{\Delta x}{2M}$ (ν_{num} is the numerical viscosity). Let us note that in 3D, the diffusive term (62) is defined with

$$B_\kappa q = \mathcal{K}\begin{pmatrix} \Delta r \\ \frac{\partial^2 u_1}{\partial x_1^2} \\ \frac{\partial^2 u_2}{\partial x_2^2} \\ \frac{\partial^2 u_3}{\partial x_3^2} \end{pmatrix} \quad \text{and} \quad \mathcal{K} = \begin{pmatrix} v_r & 0 & 0 & 0 \\ 0 & v_{u_1} & 0 & 0 \\ 0 & 0 & v_{u_2} & 0 \\ 0 & 0 & 0 & v_{u_3} \end{pmatrix} \tag{63}$$

(Δr in (62) and (63) is the Laplacian operator applied to r, and Δx in v_{num} is a strictly positive constant). In the sequel, we define the velocity viscosity $v_u := (v_{u_1}, \ldots, v_{u_d}) \in \mathbb{R}^d$ ($d \in \{1,2,3\}$). An interesting question is to verify if estimate (59) is still valid when linear wave equation (58) is replaced by first-order modified Eq. (61) whose diffusive term (62) or (63) is non-isotropic *a priori*.

4.2. Influence of the non-isotropic diffusive term

We have the following result:

Proposition 4.1. *Let $q(t,x)$ be solution of*

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = B_\kappa q, \\ q(t=0,x) = q^0(x) \end{cases} \tag{64}$$

on $\Omega = \mathbb{T}^{d\in\{1,2,3\}}$. Then:

(1) when $\Omega = \mathbb{T}^{d=1} : \forall \mathcal{K} \geqslant 0$, we have

$$\|q - \mathbb{P}q\|(t=0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M); \tag{65}$$

(2) when $\Omega = \mathbb{T}^{d\in\{2,3\}}$ and $\Delta x \gg \mathcal{O}(M) : \forall \mathcal{K} \geqslant 0$ such that $|v_u| = \mathcal{O}\left(\frac{a_* \Delta x}{M}\right)$, we have

$$\exists q^0 \in (L^2(\mathbb{T}^d))^{1+d} \quad \text{such that} \quad \|q - \mathbb{P}q\|(t=0) = \mathcal{O}(M) \quad \text{and} \quad \|q - \mathbb{P}q\|(\tau_{ac}) \gg \mathcal{O}(M) \tag{66}$$

where $\tau_{ac} = \mathcal{O}\left(\frac{M L_{\mathbb{T}^d}}{a_*}\right)$ ($L_{\mathbb{T}^d}$ is a characteristic length of $\mathbb{T}^d$ of order one and τ_{ac} is an acoustic time scale). Nevertheless, when $\Delta x = \mathcal{O}(M)$, $||q - \mathbb{P}q||(\tau_{ac}) = \mathcal{O}(M)$.

To summarize, estimate (65) is satisfied in 1D, and is not satisfied in 2D and in 3D when $\Delta x \gg \mathcal{O}(M)$ and when $|v_u| = \mathcal{O}(\frac{a_* \Delta x}{M})$ because of the loss of invariance of the well-prepared subspace $\mathcal{E}$ (see point 2 of Lemma 4.2). This loss of invariance is due to the fact that B_κ is a non-isotropic operator in 2D (and 3D). This result explains clearly (at least when the mesh is Cartesian) why 2D (and 3D) Godunov type schemes are not accurate at low Mach number in the non-linear case and why *it is the numerical viscosity on momentum equation (1)(b) that is responsible for this inaccuracy, and not the numerical viscosity on mass and energy equations (1)(a,c)*. Let us note that relation (68) shows that the acoustic energy $E_{ac}(t)$ in the acoustic subspace $\mathcal{E}^\perp$ goes to zero when $t = \mathcal{O}(1)$, $v_{u_1} = v_{u_2} = \frac{a_* \Delta x}{2M}$ and $M \ll 1$. Nevertheless, when the domain Ω is not periodic, boundary conditions can continuously inject energy in $\mathcal{E}$. In that case, the energy transfer from $\mathcal{E}$ to $\mathcal{E}^\perp$ may not converge to zero and $E_{ac}(t)$ may be of order 1 when $t = \mathcal{O}(1)$. *These statements underline that the inaccuracy of non-linear Godunov type schemes at low Mach number can be explained at least partly with linear arguments and without the convection operator $\mathcal{H}$ in (11) i.e. without the convection operator $u \cdot \nabla$ in (1), and, thus, without the notion of shear waves*. Nevertheless, non-linearities can amplify this inaccuracy (see also the discussion about non-linear system (84) that follows Conjecture 5.1). Moreover, the boundary conditions can also contribute to the loss of invariance of the well-prepared subspace $\mathcal{E}$ and, thus, to the inaccuracy at low Mach number (see Section 7.2.2).

The proof of Proposition 4.1 uses two lemmas that correspond respectively to Lemmas 2.2 and 2.3 adapted to the case of linear equation (64). The first one is given by:

Lemma 4.1. *Let $q(t,x)$ be solution of (64) on $\Omega = \mathbb{T}^{d\in\{1,2,3\}}$. Then:*

$$\forall \mathcal{K} \geqslant 0 : \quad E(t \geqslant 0) \leqslant E(t=0).$$

The second one is given by:

Lemma 4.2. *For any $\mathcal{K}$, we have:*

(1) when $\Omega = \mathbb{T}^{d=1}$: the subspaces $\mathcal{E}$ and $\mathcal{E}^\perp$ are invariant for Eq. (64);
(2) when $\Omega = \mathbb{T}^{d\in\{2,3\}}$: the subspace $\mathcal{E}$ is invariant for Eq. (64) if and only if $v_u = 0$;
(3) when $\Omega = \mathbb{T}^{d\in\{2,3\}}$: the subspace $\mathcal{E}^\perp$ is invariant for Eq. (64) if and only if $v_u = 0$.

Let us underline that the important point of Lemma 4.2 concerns the loss of invariance of $\mathcal{E}$ in 2D or 3D when $v_u \neq 0$. The fact that $\mathcal{E}^\perp$ is not invariant when $v_u \neq 0$ is not an important point (see Section 2.4). Lemma 4.2 shows also that the 1D-case and the 2D(or 3D)-case are very different: this difference is due to the fact that, when $v_u \neq 0$, the velocity diffusive term in (62) (or (63)) is isotropic if and only if the space dimension is equal to one. Let us also note that in 1D, $\mathcal{E}$ is a trivial subspace and $\mathcal{E}^\perp$ is almost equal to $(L^2(\mathbb{T}))^2$. Indeed, we have in 1D

$$\mathcal{E} = \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \in (L^2(\mathbb{T}))^2 \text{ such that } r(x) = C^{st} \text{ and } u(x) = C^{st} \right\}.$$

Of course, this is not at all the case in 2D or 3D. This difference between the 1D-case and the 2D(or 3D)-case is also underlined by the structure of the kernel $Ker(\frac{L}{M} - B_\kappa)$ with respect to v_u:

Lemma 4.3. *For any $\mathcal{K} \geqslant 0$, we have:*

(1) when $\Omega = \mathbb{T}^{d=1}$:

$$Ker\left(\frac{L}{M} - B_\kappa\right) = \mathcal{E};$$

(2) when $\Omega = \mathbb{T}^{d\in\{2,3\}}$ and $v_u = 0$:

$$Ker\left(\frac{L}{M} - B_\kappa\right) = \mathcal{E};$$

(3) when $\Omega = \mathbb{T}^{d\in\{2,3\}}$ and $v_{u_k} > 0 (k \in \{1, \ldots, d\})$:

$$Ker\left(\frac{L}{M} - B_\kappa\right) = \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \in (L^2(\mathbb{T}))^{1+d} \text{ such that } \nabla r = 0 \text{ and } \partial_{x_k} u_k = 0 \right\} \subsetneq \mathcal{E}.$$

Proof of Proposition 4.1:

- **When** $\Omega = \mathbb{T}^{d=1}$: Using Lemmas 4.1 and 4.2, point 1 is a direct consequence of point 2 of Theorem 2.2.
- **When** $\Omega = \mathbb{T}^{d\in\{2,3\}}$ **and** $v_u \neq 0$: The subspace $\mathcal{E}$ is not invariant (see point 2 of Lemma 4.2). Thus, we cannot apply point 2 of Theorem 2.2. Let us choose $q^0 \in \mathcal{E}$ such that the kth Fourier component $\check{q}^0_k$ of q^0 is given by

$$\check{q}^0_k = \frac{A_k}{\sqrt{k_1^2 + k_2^2}} \begin{pmatrix} 0 \\ -k_2 \\ k_1 \end{pmatrix} \in \check{\mathcal{E}}_{k\neq 0} \tag{67}$$

(see proof of Lemma 4.2 for the definition of $\check{\mathcal{E}}_{k\neq 0}$) with $A_k \in \mathbb{C}$, $k := (k_1, k_2)$, $k_1 := \mu_1 / L_{\mathbb{T}^d} \in \mathbb{R}$ and $k_2 := \mu_2 / L_{\mathbb{T}^d} \in \mathbb{R}$ knowing that $\mathcal{O}(|A_k|) = \mathcal{O}(|\mu_1|) = \mathcal{O}(|\mu_2|) = \mathcal{O}(L_{\mathbb{T}^d}) = 1 (L_{\mathbb{T}^d}$ is a characteristic length of $\mathbb{T}^d)$. By using (73), we obtain that $\check{q}_k(t) = \frac{A_k}{\sqrt{k_1^2+k_2^2}} \left(0, -k_2 \exp\left(-v_{u_1} k_1^2 t\right), k_1 \exp\left(-v_{u_2} k_2^2 t\right)\right)^T$. Thus, we have

$$\check{q}^\perp_k(t) = \frac{A_k}{\sqrt{k_1^2 + k_2^2}} \cdot \frac{k_1 k_2}{k_1^2 + k_2^2} \left[\exp\left(-v_{u_2} k_2^2 t\right) - \exp\left(-v_{u_1} k_1^2 t\right)\right] \begin{pmatrix} 0 \\ k_1 \\ k_2 \end{pmatrix}.$$

This implies that

$$E_{ac,k}(t) = |A_k|^2 \frac{k_1^2 k_2^2}{\left(k_1^2 + k_2^2\right)^2} \left[\exp\left(-v_{u_1} k_1^2 t\right) - \exp\left(-v_{u_2} k_2^2 t\right)\right]^2 \tag{68}$$

where $E_{ac,k}(t)$ is the kth Fourier component of the energy in the acoustic space $\mathcal{E}^\perp$. Let us choose $t = \tau_{ac} := \frac{M L_{\mathbb{T}^d}}{a_*}$ (τ_{ac} is an acoustic time scale) and let us suppose that $|v_u| = \mathcal{O}(\frac{a_* \Delta x}{M})$ which is equivalent to write that $v_{u_1} = \lambda_1 \frac{a_* \Delta x}{M}$ and $v_{u_2} = \lambda_2 \frac{a_* \Delta x}{M}$ with $\mathcal{O}(\lambda_1) + \mathcal{O}(\lambda_2) = 1 (\lambda_1 \geqslant 0$ and $\lambda_2 \geqslant 0)$. We have $v_{u_1} k_1^2 \tau_{ac} = \lambda_1 \mu_1^2 \frac{\Delta x}{L_{\mathbb{T}^d}}$ and $v_{u_2} k_2^2 \tau_{ac} = \lambda_2 \mu_2^2 \frac{\Delta x}{L_{\mathbb{T}^d}}$. This implies that

$$E_{ac,k}(\tau_{ac}) = |A_k|^2 \frac{\mu_1^2 \mu_2^2}{(\mu_1^2 + \mu_2^2)^2} \left[\exp\left(-\lambda_1 \mu_1^2 \frac{\Delta x}{L_{\mathbb{T}^d}}\right) - \exp\left(-\lambda_2 \mu_2^2 \frac{\Delta x}{L_{\mathbb{T}^d}}\right)\right]^2. \tag{69}$$

Equality (69) shows that it is possible to choose (μ_1, μ_2) of order one such that $E_{ac,k}(\tau_{ac}) = \mathcal{O}(\Delta x^2)$. By using the Parseval's theorem, we obtain the estimate $E_{ac}(\tau_{ac}) = \mathcal{O}(\Delta x^2)$ that is to say $\|q - \mathbb{P}q\|(\tau_{ac}) = \mathcal{O}(\Delta x)$ although $\|q - \mathbb{P}q\|(t = 0) = 0$. Let us now define $\tilde{q}^0 \in \mathcal{E}$ such that $\check{\tilde{q}}^0_k$ is given by (67) with $|\widetilde{A}_k| = \mathcal{O}(1)$: thus, by construction, we have $\|\tilde{q} - \mathbb{P}\tilde{q}\|(\tau_{ac}) = \mathcal{O}(\Delta x)$ where $\tilde{q}$ is solution of (64) with the initial condition $\tilde{q}^0$. Let us also define $\bar{q}^0 \in \mathcal{E}^\perp$ such that $\|\bar{q}^0\| = \mathcal{O}(M)$ and let us define $q^0 = \tilde{q}^0 + \bar{q}^0$. Then, we have $\|q^0 - \mathbb{P}q^0\| = \|\bar{q}^0\| = \mathcal{O}(M)$. Moreover, by linearity, we have $q = \tilde{q} + \bar{q}$ where $\bar{q}$ is solution of (64) with the initial condition $\bar{q}^0$. Since $\|\bar{q}^0\| = \mathcal{O}(M)$, we have also $\|\bar{q} - \mathbb{P}\bar{q}\|(\tau_{ac}) = \mathcal{O}(M)$. Thus, we obtain $\|q - \mathbb{P}q\|(\tau_{ac}) = \mathcal{O}(\Delta x)$ since $\|\tilde{q} - \mathbb{P}\tilde{q}\|(\tau_{ac}) = \mathcal{O}(\Delta x)$. □

It remains to prove Lemmas 4.1 and 4.2:

Proof of Lemma 4.1: We have $\frac{d}{dt}\|q\|^2(t) = 2\langle q, B_\kappa q\rangle$ since $\langle q, \frac{L}{M}q\rangle = \frac{a_*}{M}(\langle r, \nabla\cdot u\rangle + \langle \nabla r, u\rangle) = 0$. Thus, we obtain the result by noting that

$$\langle q, B_\kappa q\rangle = -\left(v_r\|\nabla r\|^2 + \sum_{k=1}^{d} v_{u_k}\|\partial_{x_k}u_k\|^2\right). \qquad \square \tag{70}$$

Proof of Lemma 4.2: We study the 2D-case (the 3D-case is similar). As $\mathcal{E}$ is invariant for the linear equation $\partial_t q + \frac{L}{M}q = 0$ (see Lemma 2.3 with $H = 0$), we just study the invariance of $\mathcal{E}$ for the equation

$$\begin{cases} \partial_t q = B_\kappa q, \\ q(t=0,x) = q^0(x). \end{cases} \tag{71}$$

In the Fourier space, PDE (71) is given by the ODE

$$\begin{cases} \dfrac{d}{dt}\check{q}_k = \check{B}_{\kappa,k}\check{q}_k, \\ \check{q}_k(t=0) = \check{q}_k^0 \end{cases} \tag{72}$$

where $\check{q}_k(t)$ is the kth Fourier component of $q(t,x)$ and where

$$\check{B}_{\kappa,k} = -\begin{pmatrix} v_r\|k\|^2 & 0 & 0 \\ 0 & v_{u_1}k_1^2 & 0 \\ 0 & 0 & v_{u_2}k_2^2 \end{pmatrix}.$$

In (72), $k := (k_1, k_2)$ is the wave number that belongs to an infinite discrete set $\mathbb{K}$. The solution of (72) is given by

$$\check{q}_k(t) = \begin{pmatrix} \check{r}_k^0 \exp(-v_r\|k\|^2 t) \\ \check{u}_{1,k}^0 \exp(-v_{u_1}k_1^2 t) \\ \check{u}_{2,k}^0 \exp(-v_{u_2}k_2^2 t) \end{pmatrix} \tag{73}$$

knowing that $\check{q}_k^0 = \begin{pmatrix} \check{r}_k^0 \\ \check{u}_{1,k}^0 \\ \check{u}_{2,k}^0 \end{pmatrix}$. Moreover, $q \in \mathcal{E}$ if and only if $\forall k \in \mathbb{K} : \check{q}_k \in \check{\mathcal{E}}_k$ where $\check{\mathcal{E}}_{k=0} := Vect\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$ and $\check{\mathcal{E}}_{k\neq 0} := Vect\left\{\begin{pmatrix}0\\-k_2\\k_1\end{pmatrix}\right\}$. Formula (73) shows that it is possible to find $k \in \mathbb{K}$ and $\check{q}_k^0 \in \check{\mathcal{E}}_k$ such that $\check{q}_k(t>0) \notin \check{\mathcal{E}}_k$ when $v_u \neq 0$. This proves that $\mathcal{E}$ is not invariant in 2D when $v_u \neq 0$. Of course, when $v_u = 0$, formula (73) shows that $\mathcal{E}$ is invariant in 2D. By noting that $q \in \mathcal{E}^\perp$ if and only if $\forall k \in \mathbb{K} : \check{q}_k \in \check{\mathcal{E}}_k^\perp$ where $\check{\mathcal{E}}_{k=0}^\perp := Vect\{0\}$ and $\check{\mathcal{E}}_{k\neq 0}^\perp := Vect\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\k_1\\k_2\end{pmatrix}\right\}$, we show in the same way that $\mathcal{E}^\perp$ is invariant in 2D if and only if $v_u = 0$. Let us now study the 1D-case. In that case, formula (73) is replaced by

$$\check{q}_k(t) = \begin{pmatrix} \check{r}_k^0 \exp(-v_r\|k\|^2 t) \\ \check{u}_{1,k}^0 \exp(-v_{u_1}k_1^2 t) \end{pmatrix}. \tag{74}$$

Moreover, we have $\check{\mathcal{E}}_{k=0} = \check{\mathcal{E}}_{k\neq 0}^\perp = Vect\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$ and $\check{\mathcal{E}}_{k\neq 0} = \check{\mathcal{E}}_{k=0}^\perp = \{0\}$. These relations allow to obtain that $\mathcal{E}$ and $\mathcal{E}^\perp$ are invariant in 1D for any $(v_r, v_u) \in \mathbb{R}^2$. $\square$

Proof of Lemma 4.3: When $\mathcal{K} = 0$, we have $B_\kappa = 0$. Then, we have $Ker(\frac{L}{M} - B_\kappa) = \mathcal{E}$ in any spatial dimension since $KerL = \mathcal{E}$. We now suppose that $\mathcal{K} > 0$.

- **When** $\Omega = \mathbb{T}^{d=1}$: In 1D, we have $q \in \mathcal{E} \Rightarrow r = C^{st}$ and $u = C^{st}$. Thus $Lq = B_\kappa q = 0$ that is to say $\mathcal{E} \subset Ker(\frac{L}{M} - B_\kappa)$. Let us choose $q \in Ker(\frac{L}{M} - B_\kappa)$. When $v_r > 0$ and $v_u > 0$, by using the fact that $\langle q, (\frac{L}{M} - B_\kappa)q\rangle = -\langle q, B_\kappa q\rangle$ and relation (70), we obtain that $r = C^{st}$ and $u = C^{st}$ that is to say $q \in \mathcal{E}$. When $v_r > 0$ and $v_u = 0$, we just obtain with (70) that $r = C^{st}$. Nevertheless, $r = C^{st}$ and $q \in Ker(\frac{L}{M} - B_\kappa)$ imply that $u = C^{st}$. Thus, we have also $q \in \mathcal{E}$. The case $v_r = 0$ and $v_u > 0$ is identical in 1D to the case $v_r > 0$ and $v_u = 0$.
- **When** $\Omega = \mathbb{T}^{d\in\{2,3\}}$ **and** $v_u = 0$: The proof is identical to the proof of the 1D-case by replacing $u = C^{st}$ with $\nabla\cdot u = 0$.
- **When** $\Omega = \mathbb{T}^{d\in\{2,3\}}$ **and** $v_{u_k} > 0$ $(k \in \{1,\ldots,d\})$: Let us choose $q \in Ker(\frac{L}{M} - B_\kappa)$. We deduce from (70) that $\nabla r = 0$ and $\partial_{x_k}u_k = 0$. Let us now suppose that q is such that $\nabla r = 0$ and $\partial_{x_k}u_k = 0$. In that case, we have $q \in \mathcal{E}$ (i.e. $q \in KerL$) and $q \in KerB_\kappa$ that is to say $q \in Ker(\frac{L}{M} - B_\kappa)$. $\square$

5. Construction of a class of colocated compressible schemes accurate at low Mach number

By using the results obtained in the linear case in Section 4, we now propose a large class of colocated schemes that may be accurate at low Mach number on any mesh. This class is named *low Mach X schemes* and consists in a simple modification of any X scheme of Godunov type ($X \in \{Roe, VFRoe, \ldots\}$) or not ($X$ = kinetic scheme [20] for example). We also show that other colocated schemes accurate at low Mach number as, for example, the Roe–Turkel type schemes [7–9] or the AUSM$^+$-up scheme [12] are similar to a *low Mach X scheme*. We also underline that there exists a link between the proposed *low Mach Godunov scheme* and a colocated incompressible scheme stabilized with a Brezzi–Pitkäranta type stabilization [21].

5.1. Construction of a linear low Mach Godunov scheme

It is the loss of invariance in 2D and 3D of the well-prepared subspace $\mathcal{E}$ that is responsible for inaccuracy (66) of Godunov scheme (60) applied to linear wave equation (58). This loss of invariance is induced by the fact that the velocity diffusive term in (62) and (63) is non-isotropic when $v_u \neq 0$. Thus, following point 2 of Lemma 4.2, it is natural to choose $v_u = 0$ in (62) and (63) to recover the invariance of $\mathcal{E}$ in 2D and 3D, and, finally, to recover estimate (59) in 2D and 3D by using point 2 of Theorem 2.2. More precisely, we have:

Theorem 5.1. *Let $q(t,x)$ be solution of*

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = v_r \begin{pmatrix} \Delta r \\ 0 \end{pmatrix}, \\ q(t=0,x) = q^0(x) \end{cases} \tag{75}$$

on $\Omega = \mathbb{T}^{d\in\{1,2,3\}}$. Then, $\mathbb{P}q(t \geqslant 0) = \mathbb{P}q^0$ and

$$\forall q^0 \in (L^2(\mathbb{T}^d))^{1+d},\ \forall v_r \geqslant 0 : \quad \begin{cases} E_{inc}(t \geqslant 0) = E_{inc}(t=0), \\ E_{ac}(t \geqslant 0) \leqslant E_{ac}(t=0). \end{cases}$$

This implies that:

$$\forall v_r \geqslant 0 : \quad \|q - \mathbb{P}q\|(t=0) = \mathcal{O}(M) \quad \Rightarrow \quad \|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M). \tag{76}$$

Let us note that, because of the pressure diffusive term in (75), we have $\|q - \mathbb{P}q\|(t \geqslant 0) \leqslant \|q - \mathbb{P}q\|(t=0)$. Thus, $\|q - \mathbb{P}q\|(t \geqslant 0) = \mathcal{O}(M)$ could be replaced by $\max_{t\geqslant 0}\|q - \mathbb{P}q\|(t) = \mathcal{O}(M)$ in (76). Of course, the numerical scheme, whose first-order modified equation is given by (75) (when $\forall k : \Delta x_k = \Delta x$), is the following:

$$\begin{cases} \frac{d}{dt} r_{i,j} + \dfrac{a_*}{M} \cdot \left(\dfrac{u_{1,i+1,j} - u_{1,i-1,j}}{2\Delta x_1} + \dfrac{u_{2,i,j+1} - u_{2,i,j-1}}{2\Delta x_2} \right) = \dfrac{a_*}{2M} \cdot \left(\dfrac{r_{i+1,j} - 2r_{i,j} + r_{i-1,j}}{\Delta x_1} + \dfrac{r_{i,j+1} - 2r_{i,j} + r_{i,j-1}}{\Delta x_2} \right), \\ \frac{d}{dt} u_{1,i,j} + \dfrac{a_*}{M} \cdot \dfrac{r_{i+1,j} - r_{i-1,j}}{2\Delta x_1} = 0, \\ \frac{d}{dt} u_{2,i,j} + \dfrac{a_*}{M} \cdot \dfrac{r_{i,j+1} - r_{i,j-1}}{2\Delta x_2} = 0. \end{cases} \tag{77}$$

For the sake of simplicity, linear scheme (77) is named *low Mach Godunov scheme*. In Section 5.2, we prove in the 1D-case that the explicit scheme deduced from the explicit discretization of (77) satisfies the von Neumann stability condition under a classical CFL criteria although there is numerical viscosity only on the pressure equation. We also prove in Section 5.2 that the 1D implicit scheme satisfies the von Neumann stability condition without any CFL criteria. At last, we emphasize that scheme (77) satisfies the two additional conditions introduced in Section 2.5.

5.2. Stability analysis of the 1D linear low Mach Godunov scheme

We now prove that scheme (77) written in 1D satisfies the von Neumann necessary condition for L^2 stability [27] when we discretize the time derivatives with an explicit or implicit first-order time scheme. For the explicit scheme, we have the following result:

Lemma 5.1. *Let us define the explicit scheme*

$$\begin{cases} \dfrac{r_i^{n+1} - r_i^n}{\Delta t} + \dfrac{a_*}{M} \cdot \dfrac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = \kappa_r \dfrac{a_*}{2M} \cdot \dfrac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x}, \\ \dfrac{u_i^{n+1} - u_i^n}{\Delta t} + \dfrac{a_*}{M} \cdot \dfrac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = \kappa_u \dfrac{a_*}{2M} \cdot \dfrac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x} \end{cases} \tag{78}$$

with periodic boundary conditions. Then:

(1) when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = \kappa_r$: the von Neumann condition is satisfied under the CFL condition

$$\Delta t \leqslant M \min(\kappa_r, \kappa_r^{-1}) \frac{\Delta x}{a_*}; \tag{79}$$

(2) when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = 0$: the von Neumann condition is satisfied under the CFL condition

$$\Delta t \leqslant M \min\left(\frac{\kappa_r}{2}, \kappa_r^{-1}\right) \frac{\Delta x}{a_*}. \tag{80}$$

Let us note that CFL condition (79) is optimal when $\kappa_r = 1$: this corresponds to the linear Godunov scheme. For the implicit scheme, we obtain the following result:

Lemma 5.2. *Let us define the implicit scheme*

$$\begin{cases} \dfrac{r_i^{n+1} - r_i^n}{\Delta t} + \dfrac{a_*}{M} \cdot \dfrac{u_{i+1}^{n+1} - u_{i-1}^{n+1}}{2\Delta x} = \kappa_r \dfrac{a_*}{2M} \cdot \dfrac{r_{i+1}^{n+1} - 2r_i^{n+1} + r_{i-1}^{n+1}}{\Delta x}, \\ \dfrac{u_i^{n+1} - u_i^n}{\Delta t} + \dfrac{a_*}{M} \cdot \dfrac{r_{i+1}^{n+1} - r_{i-1}^{n+1}}{2\Delta x} = \kappa_u \dfrac{a_*}{2M} \cdot \dfrac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x} \end{cases} \tag{81}$$

with periodic boundary conditions. Then, for any $\Delta t > 0$, when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = \kappa_r$ or $\kappa_u = 0$, the von Neumann condition is satisfied.

The proof of Lemmas 5.1 and 5.2 are given in Appendix B. These lemmas show that the numerical viscosity on the velocity equation should not be essential to preserve the L^2 stability as soon as the numerical viscosity on the pressure equation is strictly positive: the numerical results proposed in Section 7 justify this assertion. We emphasize that the von Neumann condition is not sufficient *a priori* to obtain the L^2 stability (see the Kreiss matrix theorem in [27], p. 74). Nevertheless, it is possible to prove that the von Neumann condition is in fact necessary and sufficient for the explicit and implicit schemes when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = \kappa_r$ and, thus, for the Godunov scheme (see the remark at the end of Appendix B). But, when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = 0$ and, thus, for the *low Mach Godunov scheme*, it seems to be more difficult to obtain a similar result. Nevertheless, in the 1D semi-discrete case, it is possible to obtain the L^2 stability when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = 0$ with an energy method [28].

5.3. Construction of non-linear colocated low Mach schemes on any mesh

The arguments used to build linear scheme (77) on a Cartesian mesh incite us to write that a good colocated non-linear X scheme of Godunov type (X = Roe [1] for example) or not (X = kinetic scheme [20] for example) is one for which the first-order modified equation associated to the linearized X scheme without convection preserves the well-prepared subspace $\mathcal{E}$. We see in Theorem 5.1 that a simple way to obtain such a property on a Cartesian mesh is to center the discretization of the pressure gradient. Moreover, the stability results of Lemmas 5.1 and 5.2 do not prohibit to apply this idea. Thus, we propose the following Conjecture for the non-linear case on any mesh:

Conjecture 5.1. *Let X be a colocated scheme of Godunov type or not applied to compressible Euler system (1) or to barotropic Euler system (6) on any 2D (or 3D) mesh. We suppose that the X scheme is stable at low Mach number. Let us modify the X scheme in using the central difference:*

(1) to discretize the momentum flux;
(2) or to discretize only the pressure gradient in the momentum flux when it is possible (X = VFRoe [2,3,8,9] or X = FDS [29,30] for example).

Then, at low Mach number:

(i) the modified X scheme remains stable;
(ii) the modified X scheme does not create any spurious wave and is accurate at low Mach number.

This modified X scheme is named "low Mach X scheme".

Let us underline that the modification applied to a X scheme to obtain a *low Mach X scheme* is *simple* and *may be applied on any mesh* (i.e. not only on a Cartesian mesh: see the discussion in Section 5.8, and the numerical results in Sections 7.1.2 and 7.2.2). Moreover, the incompressible energy is constant when $q(t,x)$ is solution of (75). This means that a *low Mach Godunov scheme* should diffuse the incompressible energy at a time scale of order 1 (here, X is a Godunov type scheme). Finally, we can say that *low Mach Godunov schemes* proposed in Conjecture 5.1 satisfy the two additional conditions introduced in Section 2.5. Let us also underline that Conjecture 5.1 is *also* written for colocated X schemes *that are not of Godunov type*.

It is important to note that Conjecture 5.1 should also be valid by adding a pressure term of order $\mathcal{O}(M^2\Delta x)$ (see the case of the linear Roe–Turkel scheme in Section 5.5.1 and the case of the AUSM$^+$-up scheme in Section 5.5.2). Indeed, on a Cartesian mesh, the 2D scheme

$$(\nabla p)_{i,j}^{discrete} := \begin{pmatrix} \dfrac{p_{i+1,j} - p_{i-1,j}}{2\Delta x_1} \\ \dfrac{p_{i,j+1} - p_{i,j-1}}{2\Delta x_2} \end{pmatrix} + \mathcal{O}(M^2 \Delta x) \tag{82}$$

is equivalent to the 2D scheme

$$(\nabla r)_{i,j}^{discrete} := \begin{pmatrix} \dfrac{r_{i+1,j} - r_{i-1,j}}{2\Delta x_1} \\ \dfrac{r_{i,j+1} - r_{i,j-1}}{2\Delta x_2} \end{pmatrix} + \mathcal{O}(M \Delta x) \tag{83}$$

because of (18) (with $\mathcal{O}(\Delta x_1) = \mathcal{O}(\Delta x_2) := \Delta x$). And, if the central difference used to discretize ∇r in linear *low Mach Godunov scheme* (77) is replaced by scheme (83), we may think that estimate (29) remains valid for first-order modified Eq. (26) associated to (83) (at least when $t \in [0, \tau_{ac}]$ where $\tau_{ac} = \mathcal{O}(M)$). In Section 6, we try to give a possible theoretical framework to justify schemes of type (83) (and, thus, of type (82)). In the sequel, we will say that a scheme verifying (83) (and, thus, (82)) is *similar* to a *low Mach X scheme*. At last, to justify Conjecture 5.1 with theoretical arguments in the non-linear case and when the mesh is Cartesian, we should prove that when we replace non-linear system (11) with the non-linear system

$$\begin{cases} \partial_t q + \mathcal{H}(q) + \dfrac{\mathcal{L}}{M}(q) = B_\kappa q, \\ q(t=0,x) = q^0(x) \end{cases} \tag{84}$$

with $\mathcal{H}(q) := 0$ or $\mathcal{H}(q)$ given by (9)(b), with $\frac{\mathcal{L}}{M}(q)$ given by (9)(c) and with $B_\kappa q$ given by (62) or (63), Theorem 2.1 remains true in 1D but remains true in 2D or 3D if and only if the velocity numerical viscosity ν_u is equal to zero in (62) or (63). This result would be the non-linear version of Proposition 4.1 and of Theorem 5.1. We do not study this difficult non-linear problem in this paper.

5.4. Application to the non-linear Roe and VFRoe schemes

To simplify the notations, we suppose that the geometry is 2D. The *low Mach Roe scheme* on a 2D Cartesian mesh is defined with (34) and (35) by replacing in (35) the matrix

$$|A_k(U)| := \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \tag{85}$$

with

$$|\widetilde{A}_k(U)| := \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}. \tag{86}$$

In the case of the VFRoe scheme [2,3] on a 2D Cartesian mesh, the fluxes are defined with

$$\begin{cases} F_{1,i+1/2,j} = \mathcal{F}_1(U^*_{i+1/2,j}), \\ F_{2,i,j+1/2} = \mathcal{F}_2(U^*_{i,j+1/2}) \end{cases} \tag{87}$$

where $\mathcal{F}_k(U)$ ($k \in \{1,2\}$) are the fluxes of compressible Euler system (1), and where $U^*_{i+1/2,j}$ and $U^*_{i,j+1/2}$ are solutions of a linearized Riemann problem. Thus, the *low Mach VFRoe scheme* is defined by replacing (87) with

$$F_{1,i+1/2,j} = \begin{pmatrix} \rho^* u_1^* \\ \rho^* u_1^{*2} + p^{**} \\ \rho^* u_1^* u_2^* \\ (\rho^* E^* + p^*) u_1^* \end{pmatrix}_{i+1/2,j} \quad \text{and} \quad F_{2,i,j+1/2} = \begin{pmatrix} \rho^* u_2^* \\ \rho^* u_1^* u_2^* \\ \rho^* u_2^{*2} + p^{**} \\ (\rho^* E^* + p^*) u_2^* \end{pmatrix}_{i,j+1/2} \tag{88}$$

where

$$\begin{cases} p^{**}_{i+1/2,j} = \dfrac{p_{i,j} + p_{i+1,j}}{2}, \\ p^{**}_{i,j+1/2} = \dfrac{p_{i,j} + p_{i,j+1}}{2}, \end{cases}$$

the quantities $(\rho^*, u_1^*, u_2^*, E^*, p^*)_{i+1/2,j}$ and $(\rho^*, u_1^*, u_2^*, E^*, p^*)_{i,j+1/2}$ being deduced from $U^*_{i+1/2,j}$ and $U^*_{i,j+1/2}$.

The *low Mach VFRoe scheme* should be more robust than the *low Mach Roe scheme*. Indeed, the *low Mach VFRoe scheme* keeps the upwinding of the discretization of $\nabla\cdot(\rho u\otimes u)$ (see point 2 of Conjecture 5.1); at the opposite, the *low Mach Roe scheme* removes any upwinding on the discrete velocity equation (see point 1 of Conjecture 5.1).

5.5. Links between low Mach X schemes and other colocated schemes accurate at low Mach number

We now prove that colocated low Mach schemes previously proposed in the literature are similar or identical to the *low Mach X scheme* proposed in Conjecture 5.1.

5.5.1. Links with the Roe–Turkel scheme and with the Roe scheme with pressure correction in the linear case

We have the following result whose proof is in Appendix C:

Proposition 5.1. *When wave equation (58) is solved on a 2D Cartesian mesh with* $\forall k:\Delta x_k=\Delta x$*:*

(1) the first-order modified equation associated to the Roe–Turkel scheme is given by (61) with

$$B_\kappa q=B_\kappa^{Roe-Turkel}q:=a_*\frac{\Delta x}{2}\begin{pmatrix}\Delta r/M^2\\ \dfrac{\partial^2 u_1}{\partial x_1^2}\\ \dfrac{\partial^2 u_2}{\partial x_2^2}\end{pmatrix}=a_*\frac{\Delta x}{2M}\left\{\begin{pmatrix}\Delta r/M\\0\\0\end{pmatrix}+\mathcal{O}(M)\right\};\tag{89}$$

(2) the first-order modified equation associated to the Roe scheme with pressure correction is given by (61) with

$$B_\kappa q=B_\kappa^{\text{pressure corr.}}q:=a_*\frac{\Delta x}{2M}\begin{pmatrix}\Delta r\\0\\0\end{pmatrix}.\tag{90}$$

Proposition 5.1 – that may be written in 3D – justifies the Roe–Turkel scheme and the Roe scheme with pressure correction. Indeed, this proposition shows that the linear Roe scheme with pressure correction is identical to the linear *low Mach Godunov scheme*; it also shows that the linear Roe–Turkel scheme is similar to the linear *low Mach Godunov scheme* since (89) implies that the pressure gradient ∇r is discretized with (83). Moreover, the pressure numerical viscosity in (89) is of order M^{-2}: this explains why the Roe–Turkel scheme has to be implicited even when the time step is of order $\frac{M\Delta x}{a_*}$ (acoustic time scale). More precisely, (89) shows that an explicit Roe–Turkel scheme should have a time step of order $\frac{M^2\Delta x}{a_*}$: this point is justified by point 2 of Lemma 5.1 by choosing $\kappa_r=M^{-1}$ and $M\ll 1$ in CFL condition (80). At the opposite, Lemma 5.2 shows that the implicit Roe–Turkel scheme should be stable without any CFL condition. These stability results are compatible with the stability analysis of Roe–Turkel scheme applied to system (1) proposed in [31]. At last, let us underline that point 2 of Proposition 5.1 coupled to point 2 of Lemma 5.1 underlines that, contrary to what it is written in [5] (see Section 4.2 in [5]), it is *a priori* possible to propose an *explicit* Godunov type scheme accurate at low Mach number that is stable with a time step of order $\frac{M\Delta x}{a_*}$.

5.5.2. Formal links with other non-linear colocated schemes

The non-linear AUSM$^+$-up scheme [12] is a flux splitting type scheme whose 2D numerical flux written in physical dimension is given by (in the x_1 direction)

$$F_{1,i+1/2,j}=F_{1,i+1/2,j}^{\text{convection}}+\begin{pmatrix}0\\p_{i+1/2,j}\\0\\0\end{pmatrix}$$

where $F_{1,i+1/2,j}^{\text{convection}}$ is an upwind discretization of the convective flux $(\rho u_1,\rho u_1^2,\rho u_1u_2,\rho u_1\mathbf{H})^T$ ($\mathbf{H}:=E+p/\rho$ is the total enthalpy) and where $p_{i+1/2,j}$ is an estimation of the pressure at the interface $(i+1/2,j)$ of the 2D Cartesian mesh. For a subsonic flow, the AUSM$^+$-up scheme is defined with

$$p_{i+1/2,j}=\mathcal{P}^+(M_{i,j})p_{i,j}+\mathcal{P}^-(M_{i+1,j})p_{i+1,j}+p_u\tag{91}$$

(see formula (75) in [12]) knowing that

$$\begin{cases}\mathcal{P}(m)^\pm=\frac14(m\pm1)^2(2\mp m)\pm\alpha m(m^2-1)^2, & \text{(a)}\\ p_u=-K_u\mathcal{P}^+(M_{i,j})\mathcal{P}^-(M_{i+1,j})(\rho_{i,j}+\rho_{i+1,j})f_a a_{i+1/2,j}(u_{i+1,j}-u_{i,j}), & \text{(b)}\\ \alpha=\frac{3}{16}(-4+5f_a^2), & \text{(c)}\\ f_a=M(2-M) & \text{(d)}\end{cases}\tag{92}$$

where $K_u \in [0,1], M_{i,j} = \dfrac{u_{i,j}}{a_{i+1/2,j}}, M_{i+1,j} = \dfrac{u_{i+1,j}}{a_{i+1/2,j}}, a_{i+1/2,j} = \mathcal{O}\left(\dfrac{a_{i,j}+a_{i+1,j}}{2}\right)$ and $M^2 = \mathcal{O}\left(\dfrac{M_{i,j}^2 + M_{i+1,j}^2}{2}\right)$. When $p_u := 0$ in (91) and $\alpha := 3/16$ in (92)(a), the AUSM$^+$-up scheme is the AUSM$^+$ scheme tested in [12,32]; when $p_u := 0$ in (91) and $\alpha := 0$ in (92)(a), the AUSM$^+$-up scheme is the AUSM scheme proposed in [33]. By using (92)(a), it is easy to obtain that

$$\mathcal{P}(m)^{\pm} = \frac{1}{2} \pm \left(\frac{3}{4} + \alpha\right) m + \mathcal{O}(m^3). \tag{93}$$

Thus, when α is defined with (92)(c), we obtain $\mathcal{P}(m)^{\pm} = \frac{1}{2} + \mathcal{O}(mM^2) + \mathcal{O}(m^3)$ (we use $f_a^2 = \mathcal{O}(M^2)$). Since p_u defined with (92)(b) is of the order of $\mathcal{O}\left(M \dfrac{u_{i+1,j} - u_{i,j}}{a_{i+1/2,j}}\right) \times \mathcal{O}(\rho a^2)$ (we use $f_a = \mathcal{O}(M)$ and $K_u = \mathcal{O}(1)$), we obtain

$$\begin{aligned} p_{i+1/2,j} &= \frac{p_{i,j}+p_{i+1,j}}{2} + \mathcal{O}\left(M \frac{u_{i+1,j}-u_{i,j}}{a_{i+1/2,j}}\right) \cdot \mathcal{O}(\rho a^2) + \mathcal{O}(M^3) \quad \text{(a)} \\ &= \frac{p_{i,j}+p_{i+1,j}}{2} + \mathcal{O}(M^2 \Delta x) \cdot \mathcal{O}(\rho a^2) + \mathcal{O}(M^3) \quad \text{(b)} \\ &= \frac{p_{i,j}+p_{i+1,j}}{2} + \mathcal{O}(M^2 \Delta x) + \mathcal{O}(M^3) \quad \text{(c)} \end{aligned} \tag{94}$$

by using (91) (we recall that $p = \mathcal{O}(\rho a^2)$). We finally obtain

$$p_{i+1/2,j} = p_* + \mathcal{O}(M^2 \Delta x) + \mathcal{O}(M^2) \tag{95}$$

when the initial conditions are well-prepared in the sense of (5). Estimate (95) is identical to estimate (47) obtained for the 1D Roe scheme. Estimate (94)(a) shows that the AUSM$^+$-up scheme is similar to a *low Mach X scheme* since (94)(a) gives an approximation of type (82) for the pressure gradient ∇p. Nevertheless, when $\alpha = 3/16$ (AUSM$^+$ scheme) or $\alpha = 0$ (AUSM scheme) for example, formula (93) shows that $\mathcal{P}(m)^{\pm} = \frac{1}{2} \pm Cm + \mathcal{O}(m^3)$ where $C = \mathcal{O}(1)$. This implies that

$$p_{i+1/2,j} = \frac{p_{i,j}+p_{i+1,j}}{2} + \mathcal{O}(M_{i,j} - M_{i+1,j}) + \mathcal{O}(M^2 \Delta x) + \mathcal{O}(M^3)$$

instead of (94)(c) (with p_u given by (92)(b) or equal to zero). Thus, since $\mathcal{O}(M_{i,j} - M_{i+1,j}) = \mathcal{O}(M \Delta x)$ in 2D (or 3D), we obtain for the 2D (or 3D) AUSM$^+$ and AUSM schemes

$$p_{i+1/2,j} = p_* + \mathcal{O}(M \Delta x) + \mathcal{O}(M^2) \tag{96}$$

when the initial conditions are well-prepared in the sense of (5). Estimate (96) is *exactly* estimate (53)(a) obtained for the 2D Roe scheme. Estimates (95) and (96) explain respectively why at low Mach number, the AUSM$^+$-up scheme should be accurate and why the AUSM$^+$ scheme (tested in [12,32]) and the AUSM scheme should not be accurate in 2D (or 3D) (when $\Delta x \gg \mathcal{O}(M)$): see Fig. 20 in [12]. Nevertheless, the previous estimates show also that the AUSM$^+$ scheme should be accurate at low Mach number when $\alpha = -3/4 + \mathcal{O}(M)$ in (92)(a): this condition is identical to condition (54) in [12]. Of course, (92)(c) satisfies this condition.

Moreover, when the flow is subsonic and when the fluid is a perfect gas, the colocated scheme proposed in [18] is given by

$$F_{1,i+1/2,j} = \frac{F_{1,i,j} + F_{1,i+1,j}}{2} + F_{1,i+1/2,j}^{\text{num. diffusion}} \tag{97}$$

with

$$F_{1,i+1/2,j}^{\text{num. diffusion}} = -\frac{1}{2} |u_{i+1/2,j}| \begin{pmatrix} \rho_{i+1,j} - \rho_{i,j} \\ \rho_{i+1,j} u_{1,i+1,j} - \rho_{i,j} u_{1,i,j} \\ \rho_{i+1,j} u_{2,i+1,j} - \rho_{i,j} u_{2,i,j} \\ \gamma(\rho_{i+1,j} E_{i+1,j} - \rho_{i,j} E_{i,j}) \end{pmatrix} \tag{98}$$

where $\gamma = \mathcal{O}(1)$ is the adiabatic constant of perfect gases (see formulae (2), (3), (10) in [18]). This scheme is accurate at low Mach number. By noting that $\gamma E = \mathbf{H} + \mathcal{O}(M^2)$ for a perfect gas at low Mach number, we see that (97) and (98) is *exactly* a *low Mach X scheme* where the convective flux $(\rho u_1, \rho u_1^2, \rho u_1 u_2, \rho u_1 \mathbf{H})^T$ is discretized with an upwind fashion. In [16,17], $F_{1,i+1/2,j}^{\text{num.diffusion}}$ is not exactly (98) but the pressure gradient is also discretized with a central difference.

To summarize, we can say that the AUSM$^+$-up scheme [12] and the colocated schemes proposed in [16–18] are respectively similar and identical to a *low Mach X scheme*.

5.6. Two other colocated schemes that do not create any spurious wave

Let us define the 2D system (the 3D-case is similar)

$$\begin{cases} \partial_t q + \dfrac{L}{M} q = B_{\mu_1,\mu_2} q, \\ q(t=0,x) = q^0(x) \end{cases} \tag{99}$$

where

$$B_{\mu_1,\mu_2}q=\frac{\mu_1}{2M}\begin{pmatrix}\Delta r\\ \dfrac{\partial^2 u_1}{\partial x_1^2}+\dfrac{\partial^2 u_2}{\partial x_1 x_2}\\ \dfrac{\partial^2 u_1}{\partial x_1 x_2}+\dfrac{\partial^2 u_2}{\partial x_2^2}\end{pmatrix}+\frac{\mu_2}{2M}\begin{pmatrix}\Delta r\\ \dfrac{\partial^2 u_1}{\partial x_1^2}+\dfrac{\partial^2 u_1}{\partial x_2^2}\\ \dfrac{\partial^2 u_2}{\partial x_1^2}+\dfrac{\partial^2 u_2}{\partial x_2^2}\end{pmatrix}=\frac{\mu_1}{2M}\begin{pmatrix}\Delta r\\ \nabla\nabla\cdot u\end{pmatrix}+\frac{\mu_2}{2M}\begin{pmatrix}\Delta r\\ \Delta u\end{pmatrix} \tag{100}$$

with $(\mu_1,\mu_2)\in\mathbb{R}^+\times\mathbb{R}^+$. We have $\frac{d}{dt}E(t)=-\frac{\mu_1}{M}(\|\nabla r\|^2+\|\nabla\cdot u\|^2)-\frac{\mu_2}{M}(\|\nabla r\|^2+\|\nabla u\|^2)\leqslant 0$ and $\frac{d}{dt}(\|\nabla r\|^2+\|\nabla\cdot u\|)^2=-\frac{\mu_1+\mu_2}{M}(\|\Delta r\|^2+\|\nabla\nabla\cdot u\|^2)\leqslant 0$. These inequalities show that the linear operator

$$\mathcal{A}_{t,\mu_1,\mu_2}:\begin{cases}(L^2(\mathbb{T}^2))^3\to(L^2(\mathbb{T}^2))^3,\\ q^0\mapsto\mathcal{A}_{t,\mu_1,\mu_2}q^0:=q(t,\cdot)\end{cases}$$

associated to (99) and (100) is a contraction and that $\mathcal{A}_{t,\mu_1,\mu_2}\mathcal{E}\subset\mathcal{E}$. Thus, we can apply point 2 of Theorem 2.2 and (29) is satisfied. Let us choose $(\mu_1,\mu_2)=(a_*\Delta x,0)$. Since $Ker(\frac{L}{M}-B_{\mu_1,0})=\mathcal{E}$ as for first-order modified Eq. (75) (see point 2 of Lemma 4.3), it should be possible to propose another class of *low Mach Godunov schemes* that do not create any spurious wave and that do not diffuse the incompressible energy at a time scale or order M by adding to fluxes (35) an appropriate discretized version of the non-isotropic diffusive operator

$$\rho_* a_*\frac{\Delta x}{2M}\begin{pmatrix}0\\ \dfrac{\partial^2 u_2}{\partial x_1 x_2}\\ \dfrac{\partial^2 u_1}{\partial x_1 x_2}\\ 0\end{pmatrix} \tag{101}$$

to compensate the non-isotropic operator $a_*\frac{\Delta x}{2M}\left(\frac{\partial^2 u_1}{\partial x_1^2},\frac{\partial^2 u_2}{\partial x_2^2}\right)^T$ in first-order modified Eqs. (61) and (62). But, the discretization of (101) on any mesh is not natural *a priori*: thus, the first additional condition introduced in Section 2.5 is not satisfied *a priori*. Nevertheless, it should be interesting to test such a modification, at last on the linear wave equation. Let us now choose $(\mu_1,\mu_2)=(0,a_*\Delta x)$. In that case, it is easy to modify any colocated scheme by adding the discretized version of the non-isotropic diffusive operator $\rho_* a_*\frac{\Delta x}{2M}\left(0,\frac{\partial^2 u_1}{\partial x_2^2},\frac{\partial^2 u_2}{\partial x_1^2},0\right)^T$ to fluxes (35) to obtain a scheme that does not create any spurious wave. Nevertheless, the energy in the well-prepared subspace $\mathcal{E}$ will decrease with a time scale of order M, and the incompressible solution will be close to a constant at a time of order 1: thus, the second additional condition introduced in Section 2.5 will not be satisfied.

5.7. Formal link with a stabilized colocated incompressible scheme

Let us now discretize the equation

$$\begin{cases}\partial_t q+\dfrac{L}{M}q=\begin{pmatrix}0\\ f+\nu\Delta u\end{pmatrix},\\ q(t=0,x)=q^0(x)\end{cases} \tag{102}$$

where $f:=-gz$ is a constant source term (g is the gravity and z is the vertical unitary vector) and where $\nu\Delta u$ is the physical dissipation ($\nu\in\mathbb{R}_*^+$ is the kinematic viscosity). Let us discretize on a 2D or 3D Cartesian mesh the acoustic operator L/M in (102) with linear *low Mach Godunov scheme* (77), the other terms being discretized with a classical scheme. Thus, the first-order modified equation is now given by

$$\begin{cases}\partial_t q+\dfrac{L}{M}q=\begin{pmatrix}v_r\Delta r\\ f+\nu\Delta u\end{pmatrix},\\ q(t=0,x)=q^0(x)\end{cases} \tag{103}$$

where $v_r=v_{num}:=\frac{a_*\Delta x}{2M}$. We have $\frac{d}{dt}\sqrt{E(t)}\leqslant\|f\|$. Thus, the energy $E(t)$ is bounded when $f\in(L^2(\mathbb{T}^d))^{1+d}$ and when $t\in[0,\tau]$ with $\tau<+\infty$. Moreover, by using the results proposed in Section 5.6, we see that for any $v_r\geqslant 0$, Eq. (103) satisfies sufficient condition (28) when f is a constant source term. This implies that estimate (29) is valid. Thus, we can write that $r=r_*+\mathcal{O}(M)$ which implies that the dynamic pressure $\Pi:=\dfrac{r-r_*}{M/a_*}$ is of order 1 (we have also $\Pi=\dfrac{p-p_*}{\rho_* M^2}$ by using (18) with $p_*=\rho_* a_*^2\left(1+\frac{M}{a_*}r_*\right)=\mathcal{O}(\rho_* a_*^2)$). Thus, we can formally write that $d\Pi=\dfrac{a_* dr}{M}$ which allows to write that Eq. (103) is formally equivalent to

$$\begin{cases} \nabla\cdot u = \xi\Delta\Pi - \left(\dfrac{M}{a_*}\right)^2 \dfrac{\partial\Pi}{\partial t}, \quad \text{with} \quad \xi = \left(\dfrac{M}{a_*}\right)^2 \cdot v_r. \\ \frac{\partial u}{\partial t} + \nabla\Pi = f + \nu\Delta u \end{cases} \tag{104}$$

Let us now choose $v_r = \lambda\cdot\left(\frac{a_*\Delta x}{M}\right)^2$ where λ is a constant of order 1 (we recall that Eq. (103) satisfies (28) as soon as $v_r \geqslant 0$ and f is a constant source term: see above). Then, any stationary solution of (104) is solution of

$$\begin{cases} \nabla\cdot u = \lambda\Delta x^2\Delta\Pi, & \text{(a)} \\ \nabla\Pi = f + \nu\Delta u. & \text{(b)} \end{cases} \tag{105}$$

Eq. (105) is exactly the modified equation obtained in [21]. The term $\lambda\Delta x^2\Delta\Pi$ in (105)(a) is a Brezzi–Pitkäranta type stabilization [34] (see (3.1) in [34]). This term allows to control the checkerboard modes induced by a colocated discretization of the Stokes system [21]. This is completely coherent with the results proposed in [28].

5.8. Analysis on any mesh

The analysis proposed in Sections 4 and 5 explains with linear arguments the inaccuracy of Godunov type schemes at low Mach number and justifies Conjecture 5.1 only when the mesh is Cartesian. Indeed, the notion of modified equation associated to a numerical scheme is central in our analysis, and it is *a priori* difficult to define such a notion when the mesh is not Cartesian. Nevertheless, the central argument concerns the invariance of the well-prepared subspace $\mathcal{E}$ for the linear wave equation (see sufficient condition (28)). This means that, to study the problem on any mesh, we have to generalize the notion of invariance at the semi-discrete level (i.e. continuous in time and discrete in space). The first step is to define for a given mesh a discrete well-prepared subspace $\mathcal{E}_{discrete}$ that is a good approximation of the continuous well-prepared subspace $\mathcal{E}$. The second step is to study how any discrete initial condition close to $\mathcal{E}_{discrete}$ is changed by the semi-discrete linear scheme. We do not study this question in this paper.

Nevertheless, let us say that in the case of the 1D linear Godunov scheme, the work is simple. In [28], we build a discrete well-prepared subspace $\mathcal{E}_{discrete}$ that is invariant for the 1D linear Godunov scheme and stationary for the 1D linear *low Mach Godunov scheme*. The results proposed in [28] are coherent with the first point of Lemma 4.2 and with Theorem 5.1.

When the mesh is triangular (or tetrahedral) and compatible with periodic boundary conditions, it is possible to define a discrete well-prepared subspace $\mathcal{E}_{discrete}$ that is stationary (and, thus, invariant) for the Godunov scheme and for the *low Mach Godunov scheme* [35]. Let us underline that when the mesh is Cartesian, we have the same result only in the case of the *low Mach Godunov scheme* [35], this result being the discrete version of Lemma 4.3.

These properties justify the results proposed in [36,37]. Indeed, it is proven in [36,37] with a formal asymptotic analysis and with graph theory arguments that, when the initial conditions are well-prepared in the sense of (5), the 2D non-linear Godunov scheme does not create any spurious waves when the mesh is a particular triangular mesh and when the spatial domain Ω is periodic (or unbounded in each direction). In [35], this result is proven in the linear case on any triangular (or tetrahedral) mesh.

The fact that the linear or non-linear Godunov scheme with periodic boundary conditions has a behavior on a triangular or tetrahedral mesh that is not the same than the one obtained on a Cartesian mesh is another argument which emphasizes the importance of the space dimension in the analysis of the inaccuracy of Godunov type schemes at low Mach number (cf. Section 3.2 and Lemma 4.2). Indeed, contrary to a Godunov type scheme on a Cartesian mesh, a Godunov type scheme on a triangular mesh cannot be seen as a simple combination of two 1D Godunov type schemes: in a way, the numerical diffusion on the velocity equation is "more" isotropic when the mesh is triangular (or tetrahedral) (we remind that the loss of invariance of $\mathcal{E}$ in Lemma 4.2 is linked to the fact that the velocity numerical diffusion in (60) is non-isotropic).

Nevertheless, this particular property of Godunov type schemes obtained on a triangular or tetrahedral mesh could be limited although it is really remarkable. Indeed, it is obtained by supposing that the spatial domain Ω is periodic or unbounded. When it is not the case, the proof in [35] is not valid. As a consequence, Godunov type schemes on any triangular or tetrahedral mesh with non-special boundary conditions on $\partial\Omega$ could create spurious waves at low Mach number as on a Cartesian mesh. Numerical results proposed in Section 7.2 seem to confirm this assertion.

6. A weaker sufficient condition to avoid the creation of spurious waves at low Mach number

To satisfy (29), we have constructed the linear *low Mach Godunov scheme* in Section 5 in such a way the well-prepared subspace $\mathcal{E}$ is invariant for first-order modified Eq. (33) (cf. condition (28)). In fact, we can propose a weaker sufficient condition that only imposes the invariance of a family of subspaces $\mathcal{E}_{M\Delta x}$ that are not $\mathcal{E}$ when $M\Delta x \neq 0$ but that are close to $\mathcal{E}$ when $M\Delta x \ll 1$. Indeed, we can construct $\mathcal{F}_x^{low\ Mach\ X,L/M}$ in such a way, for any $(M,\Delta x) \in \mathbb{R}_*^+ \times \mathbb{R}^+$ such that $M\Delta x \ll 1$, there exists a subspace $\mathcal{E}_{M\Delta x}$ of $(L^2(\mathbb{T}^d))^{1+d}$ that satisfies

$$\begin{cases} \mathcal{E}_{M\Delta x} \oplus \mathcal{E}^\perp = (L^2(\mathbb{T}^d))^{1+d}, & \text{(a)} \\ \forall q \in \mathcal{E}_{M\Delta x}:\ \|q - \mathbb{P}q\| \leqslant M\Delta x\|q\|, & \text{(b)} \\ \forall q^0 \in \mathcal{E}_{M\Delta x}:\ q(t \geqslant 0) \in \mathcal{E}_{M\Delta x} & \text{(c)} \end{cases} \tag{106}$$

since point 2 of Theorem 2.2 is valid when (28) is replaced by weaker condition (106). This weaker sufficient condition could be a starting point to propose other modifications of the Godunov scheme than the one proposed in Conjecture 5.1 (see scheme (82) that is equivalent to scheme (83)) or to study the influence of the nth-order terms ($n > 1$) of the modified equation. To obtain (29) with (106), we use the following lemma (whose proof is in Appendix A):

Lemma 6.1. *Let $\mathcal{A}_{\mu\geqslant 0}$ be a family of linear applications on a Hilbert space $\mathbb{H}$ and let us suppose that $\mathcal{E}$ and $\mathcal{E}_\mu$ are vector subspaces of $\mathbb{H}$ such that $\mathcal{E}\oplus\mathcal{E}^\perp = \mathcal{E}_\mu\oplus\mathcal{E}^\perp = \mathbb{H}$. Let us define the Hilbertian norm $\|\cdot\|$ and another norm $|||\cdot|||$ in $\mathbb{H}$ possibly different to $\|\cdot\|$. Then, the conditions*

$$\begin{cases} \exists(\lambda,\varepsilon)\in\mathbb{R}_*^+\times\mathbb{R}_*^+ \text{such that } \forall\mu\in[0,\varepsilon[:\ \max(\|\mathcal{A}_\mu\|,|||\mathcal{A}_\mu|||)\leqslant\lambda, & \text{(a)}\\ \forall q\in\mathcal{E}_\mu:\ |||(\mathbf{1}-\mathbb{P})q\|\leqslant\mu|||q|||, & \text{(b)}\\ \mathcal{A}_\mu\mathcal{E}_\mu\subset\mathcal{E}_\mu & \text{(c)}\end{cases} \tag{107}$$

imply that

$$\forall q\in\mathbb{H},\ \forall\mu\in[0,\varepsilon[:\quad \|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|\leqslant 2\lambda\mu|||\mathbb{P}_\mu q|||+\lambda\|(\mathbf{1}-\mathbb{P})q\| \tag{108}$$

where $\mathbb{P}$ and $\mathbb{P}_\mu$ are respectively the orthogonal projection onto $\mathcal{E}$ and the projection onto $\mathcal{E}_\mu$ along $\mathcal{E}^\perp$. By supposing also that there exists $\beta>0$ such that $\forall q\in\mathbb{H}:|||q|||\leqslant\beta\|q\|$, we obtain the more precise inequality

$$\forall q\in\mathbb{H},\ \forall\mu\in[0,\min(\varepsilon,\beta^{-1})[:\quad \|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|\leqslant\frac{2\lambda\beta\mu}{1-\beta\mu}\|q\|+\lambda\|(\mathbf{1}-\mathbb{P})q\|. \tag{109}$$

This proves that when $\mu\ll 1$, we have

$$\forall q\in\mathbb{H}:\quad \|(\mathbf{1}-\mathbb{P})q\|=\mathcal{O}(M)\quad\Rightarrow\quad\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|=\mathcal{O}(\mu)+\mathcal{O}(M) \tag{110}$$

that is to say

$$\forall q\in\mathbb{H}:\quad \|(\mathbf{1}-\mathbb{P})q\|=\mathcal{O}(M)\quad\Rightarrow\quad\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|=\mathcal{O}(M)$$

when $\mu=\mathcal{O}(M)\ll 1$.

Condition (107)(a) is natural since, in the context of the paper, $\mathcal{A}_\mu$ is supposed to converge when μ goes to zero to the bounded linear operator $\mathcal{A}_t^{(0,L/M)}$ defined with (24) and associated to the linear acoustic operator L/M ($\mathcal{A}_t^{(0,L/M)}$ is an isometry for the H^1-norm $|||\cdot||| = \|\cdot\|_{H^1}$). Condition (107)(b) means that $\mathcal{E}_{\mu=0}=\mathcal{E}$ (this is a consequence of $\mathcal{E}\oplus\mathcal{E}^\perp=\mathcal{E}_\mu\oplus\mathcal{E}^\perp=\mathbb{H}$). Then, $\mathcal{E}_\mu$ can be seen as the "sum" of $\mathcal{E}$ and of a "perturbation" ϵ_μ of order "$\mathcal{O}(\mu)$" when $\mu\ll 1$. We introduce the second norm $|||\cdot|||$ because $\mathcal{E}_\mu$ may need more regularity than $\mathcal{E}$ (see the example below). In finite dimension, we do not need to introduce this second norm. Moreover, we do not suppose in Lemma 6.1 that $\mathcal{A}_\mu\mathcal{E}\subset\mathcal{E}$ when $\mu\in[0,\varepsilon[$. Indeed, if $\mathcal{A}_\mu\mathcal{E}\subset\mathcal{E}$ when $\mu\in[0,\varepsilon[$, condition (28) would be satisfied for any $\mu\in[0,\varepsilon[$ and we could apply point 2 of Theorem 2.2 to obtain (29). Let us note that condition (107)(b) is weaker than $\forall q\in\mathbb{H}:\|(\mathbb{P}_\mu-\mathbb{P})q\|\leqslant\mu|||q|||$. To summarize, since $\mathcal{E}$ is the kernel of the acoustic linear operator L/M – that is to say $\forall q\in\mathcal{E}:\mathcal{A}_t^{(0,L/M)}q=q$ –, the conditions on $\mathcal{E}_\mu$ and $\mathcal{A}_\mu$ mean that this kernel is not really pertubed – that is to say $\mathcal{E}_\mu\simeq\mathcal{E}$ and $\forall q\in\mathcal{E}_\mu:\ ||\mathcal{A}_\mu q-q||\simeq 0$ when $\mu\ll 1$ – even if zero is not an eigenvalue of $\mathcal{F}_x^{\text{low Mach }X,L/M}$ when $\Delta x\neq 0$. For example, let us define the system

$$\begin{cases}\partial_t q+\dfrac{L}{M}q=B_\mu q,\\ q(t=0,x)=q^0(x)\end{cases} \tag{111}$$

with

$$B_\mu q=\frac{\mu}{2M}\begin{pmatrix}0\\ \dfrac{\partial^2u_1}{\partial x_1^2}+2\dfrac{\partial^2u_2}{\partial x_1x_2}\\ \dfrac{\partial^2u_2}{\partial x_2^2}\end{pmatrix}. \tag{112}$$

We associate to PDE (111) the linear application $\mathcal{A}_{t,\mu}:\begin{cases}(L^2(\mathbb{T}^2))^3\to(L^2(\mathbb{T}^2))^3,\\ q^0\mapsto\mathcal{A}_{t,\mu}q^0:=q(t,\cdot).\end{cases}$. We verify that $\frac{d}{dt}E(t)=-\frac{\mu}{M}\|\nabla\cdot u\|^2\leqslant 0$. Thus, $\mathcal{A}_{t,\mu}$ is a contraction on $\mathbb{H}=(L^2(\mathbb{T}^2))^3$ for the Hilbertian norm $\|\cdot\|=\|\cdot\|_{L^2}$. In other words, we have $\|\mathcal{A}_{t,\mu}\|\leqslant 1$. The second norm $|||\cdot|||$ is defined with the norm in $(H^1(\mathbb{T}^2))^3$ and is now noted $\|\cdot\|_{H^1}$. We admit that (111) is well-posed in $L^\infty([0,+\infty[;(H^s(\mathbb{T}^2))^3)$ where $s\in\mathbb{N}$ is large enough and that $\lim_{\mu\to 0}\|\mathcal{A}_{t,\mu}-\mathcal{A}_t^{(0,L/M)}\|_{H^1}=0$. Thus, since $\|\mathcal{A}_t^{(0,L/M)}\|_{H^1}=1$, there exists $\lambda>1$ and $\varepsilon>0$ such that for any $\mu\in[0,\varepsilon[$, we have $\|\mathcal{A}_{t,\mu}\|_{H^1}\leqslant\lambda$. This proves that condition (107)(a) is satisfied. We also verify that the subspace

$$\mathcal{E}_\mu:=\left\{q:=\begin{pmatrix}r\\ u_1\\ u_2\end{pmatrix}\in(L^2(\mathbb{T}^2))^3 \text{such that } q=\hat{q}+\frac{\mu}{2a_*}\begin{pmatrix}\dfrac{\partial u_2}{\partial x_2}\\ 0\\ 0\end{pmatrix}\text{where }\hat{q}:=\begin{pmatrix}r_*\\ u_1\\ u_2\end{pmatrix}\in\mathcal{E}\right\}$$

is the kernel of the operator $\frac{L}{M} - B_\mu$. Thus, $\mathcal{A}_{t,\mu}\mathcal{E}_\mu = \mathcal{E}_\mu$ and condition (107)(c) is satisfied. Moreover, we have $\mathcal{E}_\mu \oplus \mathcal{E}^\perp = (L^2(\mathbb{T}^2))^3$ and $\hat{q} = \mathbb{P}q$ since $\hat{q} \in \mathcal{E}$ and $\left(\frac{\partial u_2}{\partial x_2}, 0, 0\right)^T \in \mathcal{E}^\perp$, and we have

$$\forall q \in \mathcal{E}_\mu : \ \|(\mathbf{1} - \mathbb{P})q\| = \frac{\mu}{2a_*}\left\|\left(\frac{\partial u_2}{\partial x_2}, 0, 0\right)^T\right\| \leqslant \frac{\mu}{2a_*}\|q\|_{H^1}$$

where $\mathbb{P}$ is the Hodge projection (see Lemma 2.1 for the definition of the Hodge decomposition). This proves that condition (107)(b) is satisfied (with μ replaced by $\mu/(2a_*)$). Then, we can apply Lemma 6.1 to obtain inequality (108) (with μ replaced by $\mu/(2a_*)$). But, since there does not exist $\beta > 0$ such that $\forall q \in (L^2(\mathbb{T}^2))^3 : \|q\|_{H^1} \leqslant \beta\|q\|$, we cannot rigorously deduce estimate (110) from (108) when $\mu \ll 1$. Nevertheless, since $\mathbb{P}_\mu q = \mathbb{P}q + \frac{\mu}{2a_*}\left(\frac{\partial u_2}{\partial x_2}, 0, 0\right)^T$, we have

$$\|\mathbb{P}_\mu q\|_{H^1} \leqslant \|\mathbb{P}q\|_{H^1} + \frac{\mu}{2a_*}\left\|\left(\frac{\partial u_2}{\partial x_2}, 0, 0\right)^T\right\|_{H^1} \leqslant \|\mathbb{P}q\|_{H^1} + \frac{\mu}{2a_*}\|q\|_{H^2}$$

when q is enough regular ($q \in (H^s(\mathbb{T}^2))^3$ with $s \geqslant 2$), $\|\cdot\|_{H^2}$ being the norm in $(H^2(\mathbb{T}^2))^3$. Thus, we can replace inequality (108) with

$$\forall q \in (H^2(\mathbb{T}^2))^3, \ \forall \mu \in [0, \varepsilon[: \quad \|(\mathbf{1} - \mathbb{P})\mathcal{A}_{t,\mu}q\| \leqslant \frac{\lambda\mu}{a_*}\left(\|\mathbb{P}q\|_{H^1} + \frac{\mu}{2a_*}\|q\|_{H^2}\right) + \lambda\|(\mathbf{1} - \mathbb{P})q\|. \tag{113}$$

We deduce from inequality (113) that estimate (29) is satisfied when $\mu = a_* M\Delta x$ and $M \in [0, \varepsilon/(a_*\Delta x)[$. To summarize, we have obtained that when numerical diffusion operator (112) is given by

$$B_{M\Delta x}q = a_*\frac{\Delta x}{2}\begin{pmatrix} 0 \\ \dfrac{\partial^2 u_1}{\partial x_1^2} + 2\dfrac{\partial^2 u_2}{\partial x_1 x_2} \\ \dfrac{\partial^2 u_2}{\partial x_2^2} \end{pmatrix}, \tag{114}$$

estimate (29) is valid for PDE (111) when $M \ll 1$ and when the initial condition q^0 is enough regular ($q^0 \in (H^s(\mathbb{T}^2))^3$ with $s \geqslant 2$). Diffusion operator (114) is similar to operator (89) obtained with the linear Roe–Turkel scheme except on the pressure equation, and it defines a scheme of type (83). Nevertheless, it is *a priori* difficult to use (114) to propose another *simple* class of *low Mach Godunov schemes* in the non-linear case. Indeed, as for the Roe–Turkel scheme (see proof of point 1 of Proposition 5.1 in Appendix C), it is difficult to find a basic modification to apply to a non-linear Godunov type scheme in such a way the first-order modified equation is given by (111) and (114) in the linear case: in other words, the first of the two additional conditions introduced in Section 2.5 is not satisfied. At the opposite, the non-linear *low Mach Godunov type schemes* proposed in Conjecture 5.1 satisfy these two additional conditions.

7. Numerical results

We show numerical results that prove that the *Low Mach X schemes* proposed in Section 5.4 (see also Conjecture 5.1) are accurate at low Mach number when the *X* scheme is a Roe scheme [1] or a VFRoe scheme [2,3]. In the first subsection, we solve the compressible Euler system in an unbounded (physical) domain (but, of course, in a bounded numerical domain); in the second subsection, we solve the compressible Navier–Stokes system in a bounded (physical and numerical) domain.

Let us underline that boundary conditions are not taken into account from a theoretical point of view in this paper although it is an important question. In particular, some numerical results seem to show that the *low Mach Roe (or VFRoe) scheme* may not be stable when the upwinding on the pressure gradient is also deleted on *all* the interfaces that belong to the boundary of the (numerical or physical) domain (we recall that the stability results proposed in Lemmas 5.1 and 5.2 are proven when the domain is periodic). Thus, to avoid such potential numerical instabilities, we do not change the Roe (or VFRoe) numerical fluxes on the interfaces that belong to the boundary of the domain: numerical results proposed in this section seem to show that this stabilizes the *low Mach Roe (or VFRoe) scheme* without changing the accuracy properties obtained in a periodic domain at low Mach number.

7.1. Compressible Euler system in an unbounded physical domain

The test-case consists in a flow in an unbounded 2D channel with a bump defined by

$$x_2^{bump}(x_1) = \frac{1}{10}\cdot[1 - \cos((x_1 - 1)\pi)] \text{ if } x_1 \in [1, 3], = 0 \text{ if } x_1 \notin [1, 3].$$

The fluid is a perfect gas that is to say $p = \rho RT$ and $h = \frac{\gamma RT}{\gamma - 1}$ where $R = 287$ J kg^{-1} K^{-1} and $\gamma = 1,4$. Although the physical domain is unbounded, the (numerical) domain Ω – where the simulation is done – is of course bounded; it is included into the domain $[0, 4] \times [0, 1]$. The unboundedness of the physical domain is taken into account through the boundary conditions on $x_1 = 0$ and $x_1 = 4$ where the flow is supposed to be defined by the flow at the infinity. The boundary condition on $x_1 = 0$ is

defined by $(p,h)=(p_0,h_0)$ and $u=u_0:=(u_{in},0)$ with $p_0=10^5$ Pa and $h_0=25$ kJ kg^{-1}. The velocity u_{in} is a strictly positive constant such that the initial Mach number $M:=\frac{\|u_0\|}{\sqrt{\gamma p_0/\rho_0}}$ is equal to 10^{-2} or 10^{-3}: here, we have the relation $u_{in}=M\times 10^2$ m s^{-1}. The boundary condition on $x_1=4$ is defined by $p=p_0$. The boundary condition on $x_2^{bump}(x_1)$ and $x_2=1$ is a slipping boundary condition. The initial condition is given by $(p,h,u)=(p_0,h_0,u_0)$. We test the *low Mach Roe scheme* and the *low Mach VFRoe scheme* proposed in Section 5.4 with a quadrangular or triangular mesh of only 40×10 cells.

7.1.1. Results on a quadrangular mesh

The iso-Mach and iso-pressure lines on Figs. 1 and 2 show that the solution given by the Roe scheme is not correct (see Fig. 11 in [38] for a correct solution). At the opposite, Figs. 3–6 and 8 show that the solutions given by the *low Mach Roe*

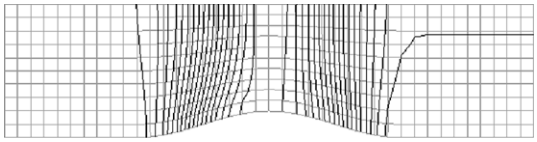

Fig. 1. Iso-Mach, Roe scheme, Mach = 10^{-2}.

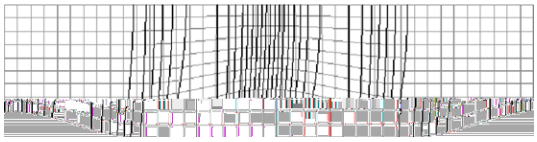

Fig. 2. Iso-pressure, Roe scheme, Mach = 10^{-2}.

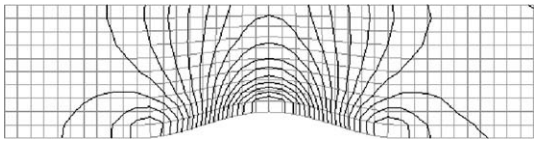

Fig. 3. Iso-Mach, *Low Mach Roe scheme*, Mach = 10^{-2}.

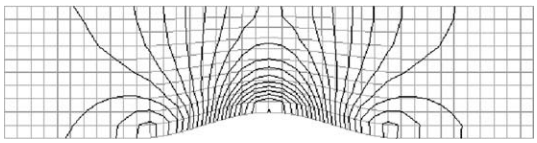

Fig. 4. Iso-pressure, *Low Mach Roe scheme*, Mach = 10^{-2}.

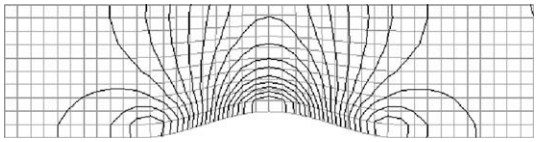

Fig. 5. Iso-Mach, *Low Mach Roe scheme*, Mach = 10^{-3}.

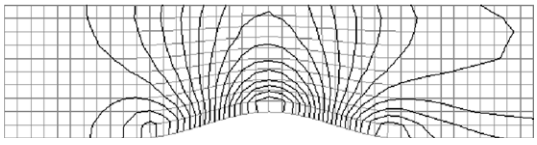

Fig. 6. Iso-Mach, *Low Mach VFRoe scheme*, Mach = 10^{-2}.

scheme and the *low Mach VFRoe scheme* are close to the incompressible solution on a quadrangular mesh. Figs. 7 and 8 show that the result is better with the *low Mach VFRoe scheme* than with the *low Mach Roe scheme* on the density field: this is perhaps due to the fact that the *low Mach Roe scheme* is built by also deleting the upwinding of the discretization of the convective operator $\nabla \cdot (\rho u \otimes u)$ which is not the case with the *low Mach VFRoe scheme* (see Conjecture 5.1 and Section 5.4). Nevertheless, the result on the density field is not really important since any $\mathcal{O}(M)$ density perturbation does not perturbe the velocity and pressure fields (see well-prepared initial conditions (5)). Let us underline that the results on Figs. 3 and 5

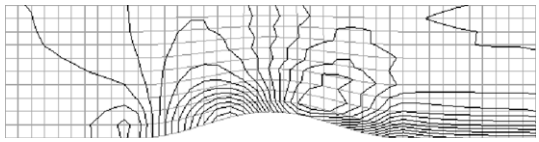

Fig. 7. Iso-density, *Low Mach Roe scheme*, Mach = 10^{-2}.

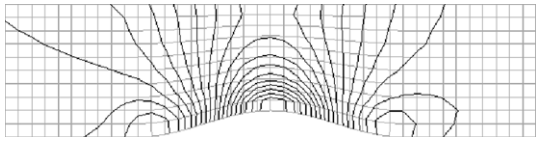

Fig. 8. Iso-density, *Low Mach VFRoe scheme*, Mach = 10^{-2}.

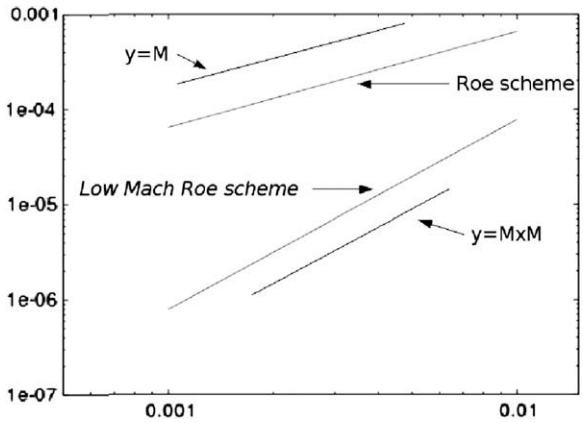

Fig. 9. $\frac{p_{max}-p_{min}}{p_{max}}(M)$, Cartesian mesh.

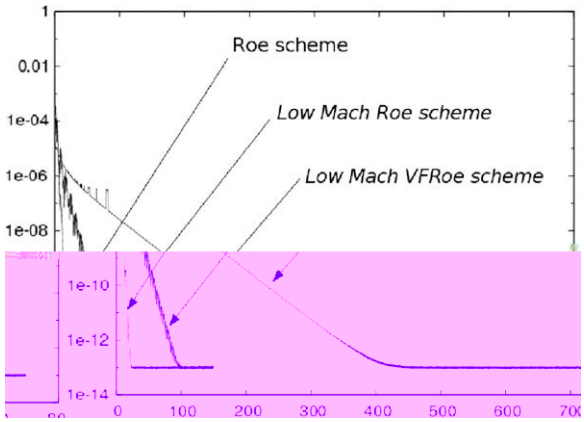

Fig. 10. Residual, Cartesian mesh, Mach = 10^{-2}.

show that the solution given by the *Low Mach Roe scheme* does not depend on the Mach number M: of course, any incompressible solution has to satisfy this property.

Fig. 9 shows that the pressure fluctuations are of order M with the Roe scheme and of order M^2 with the *low Mach Roe scheme* as with the Roe–Turkel or VFRoe–Turkel schemes [7,8]. This figure confirms that the Roe scheme creates spurious pressure waves of order $\mathcal{O}(M\Delta x)$ and that the *low Mach Roe scheme* does not create such spurious pressure waves. Figs. 10 and 11 show the residuals. Fig. 10 shows that the Roe scheme converges. Nevertheless, it converges toward a bad solution (see Figs. 1 and 2). We can see on Fig. 10 that the *low Mach VFRoe scheme* needs more iterations to converge to the stationary solution than the *low Mach Roe scheme*. Fig. 11 shows that the more the Mach number is close to zero, the more it is difficult

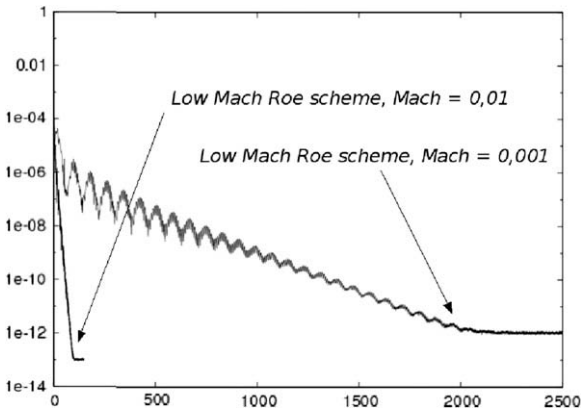

Fig. 11. Residual, Cartesian mesh, Mach = 10^{-2} and 10^{-3}.

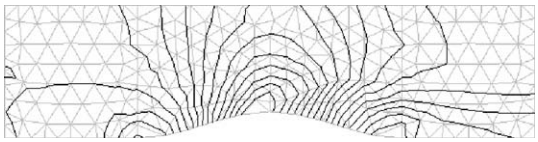

Fig. 12. Iso-Mach, *Low Mach VFRoe scheme*, Mach = 10^{-2}.

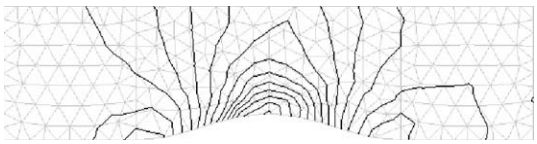

Fig. 13. Iso-pressure, *Low Mach VFRoe scheme*, Mach = 10^{-2}.

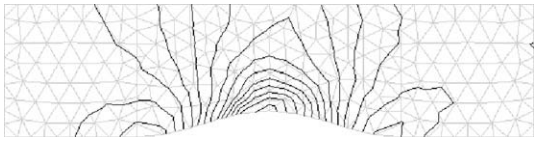

Fig. 14. Iso-density, *Low Mach VFRoe scheme*, Mach = 10^{-2}.

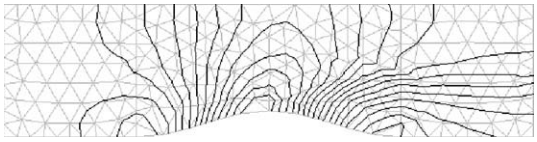

Fig. 15. Iso-Mach, VFRoe scheme, Mach = 10^{-2}.

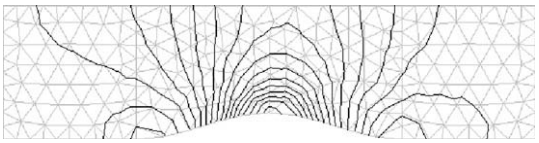

Fig. 16. Iso-pressure, VFRoe scheme, Mach = 10^{-2}.

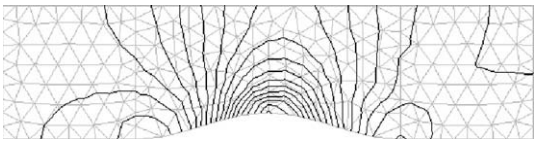

Fig. 17. Iso-density, VFRoe scheme, Mach = 10^{-2}.

to converge toward the incompressible solution. Of course, this is due to the existence of two different time scales in compressible Euler system (1) when the Mach number goes to zero.

7.1.2. Results on a triangular mesh

We have also tested the *Low Mach VFRoe scheme* on a triangular mesh when $M = 10^{-2}$. Figs. 12–14 confirm the previous results. Moreover, Figs. 15–17 confirm that Godunov type schemes do not suffer of any inaccuracy at low Mach number on triangular (or tetrahedral) mesh when the domain is periodic or unbounded [35–37] (see also Section 5.8). Fig. 18 shows that the *low Mach VFRoe scheme* on a triangular mesh needs more iterations than the VFRoe scheme to converge. We can also note that the *low Mach VFRoe scheme* converges with less iterations on a triangular mesh than on a quadrangular mesh (compare Figs. 10 and 18).

7.2. Compressible Navier–Stokes system in a bounded physical domain

We now study the 2D natural convection test-case studied in [39–41] by solving the compressible Navier–Stokes system with a gravity term ($g = 9.81\ \mathrm{m\ s^{-2}}$) in a bounded (physical) domain Ω. The domain Ω is a 2D square whose area is equal to $L \times L$ with $L = 1.528 \times 10^{-3}$ m. The fluid is the perfect gas defined in Section 7.1 with $\mu = 1.619 \times 10^{-6}\ \mathrm{kg\ m^{-1}\ s^{-1}}$ (viscosity) and $\lambda = 2.29 \times 10^{-3}\ \mathrm{W\ m^{-1}\ K^{-1}}$ (conductivity). The boundary conditions are the following:

$$\begin{cases} x_1 = 0: & T = 283.15\ \mathrm{K} \quad \text{and} \quad u = 0\ \mathrm{m\ s^{-1}}, \\ x_1 = L: & T = 263.15\ \mathrm{K} \quad \text{and} \quad u = 0\ \mathrm{m\ s^{-1}}, \\ x_2 = 0: & \partial_{x_2} T = 0\ \mathrm{K\ m^{-1}} \quad \text{and} \quad u = 0\ \mathrm{m\ s^{-1}}, \\ x_2 = L: & \partial_{x_2} T = 0\ \mathrm{K\ m^{-1}} \quad \text{and} \quad u = 0\ \mathrm{m\ s^{-1}}. \end{cases} \tag{115}$$

We can deduce from the previous physical quantities that the flow is such that $M \simeq 10^{-4}$ (Mach number), $Ra \simeq 10^3$ (Rayleigh number) and $Pr \simeq 0.71$ (Prandtl number), which gives a Reynolds number $Re := \sqrt{Ra/Pr}$ of about 37. The convective part of the system is discretized with the Roe scheme or with the *low Mach Roe scheme* proposed in Section 5.4. The physical diffusive part of the system is discretized with a finite volume type scheme.

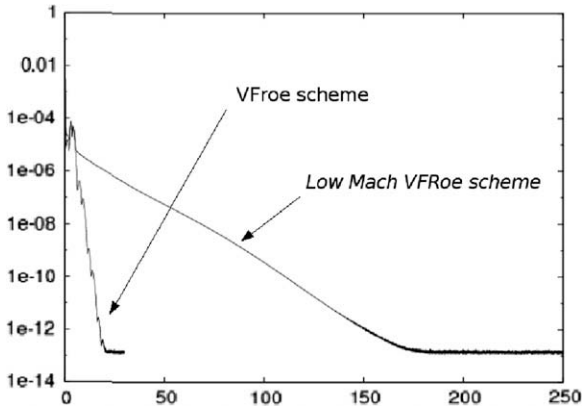

Fig. 18. Residual, Triangular mesh, Mach = 10^{-2}.

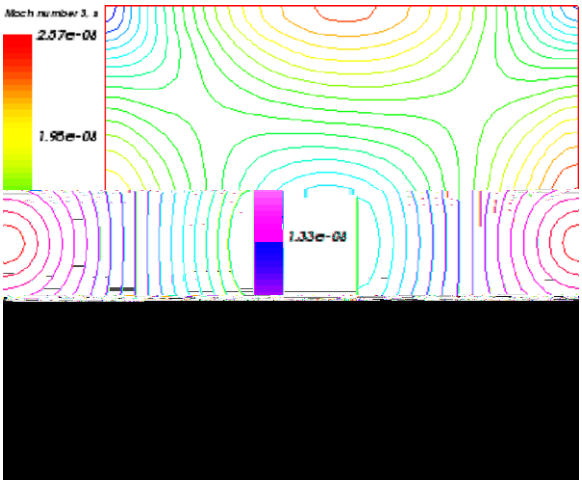

Fig. 19. Iso-Mach, Roe scheme, Cartesian mesh.

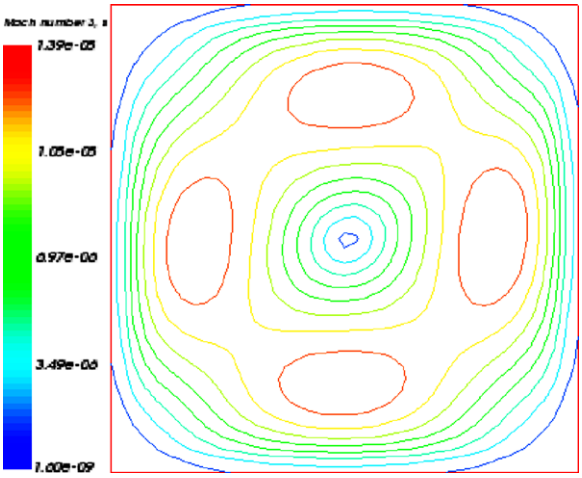

Fig. 20. Iso-Mach, *Low Mach Roe scheme*, Cartesian mesh.

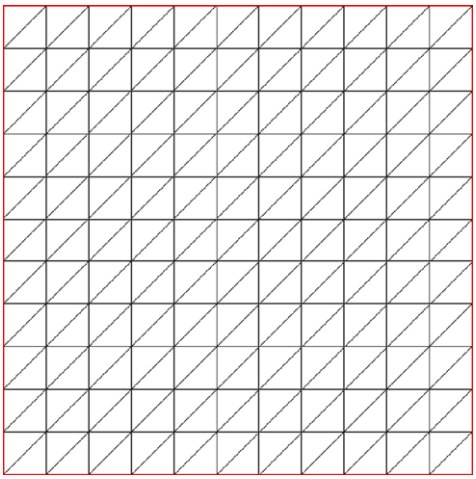

Fig. 21. 1st triangular mesh.

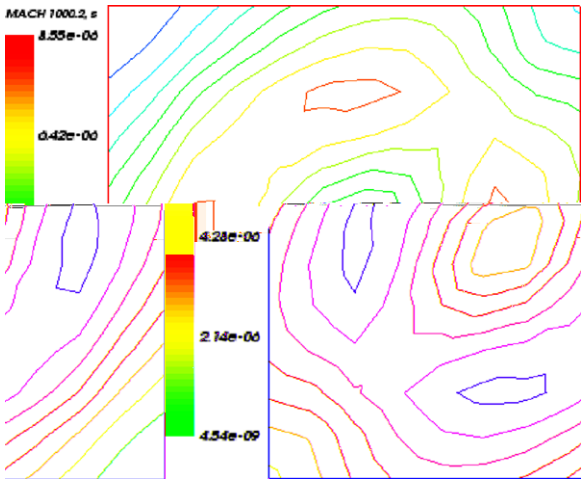

Fig. 22. Iso-Mach, Roe scheme, 1st triangular mesh.

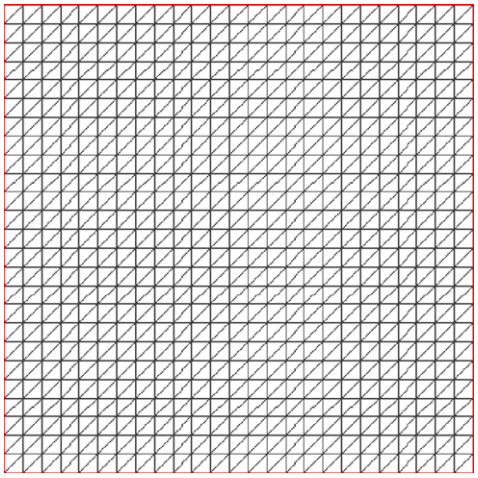

Fig. 23. 2D triangular mesh.

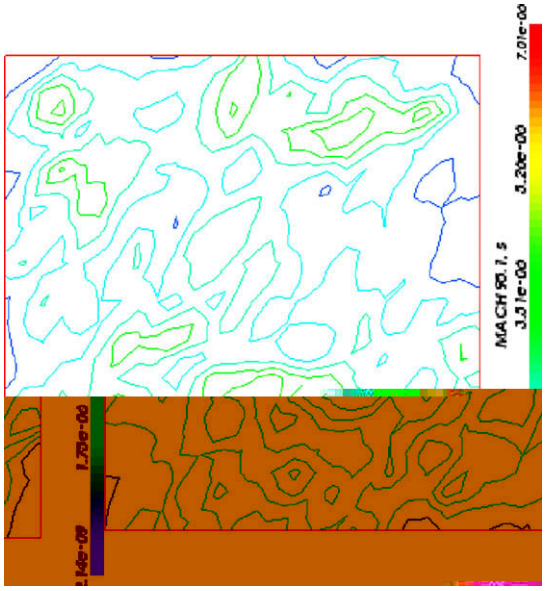

Fig. 24. Iso-Mach, Roe scheme, 2D triangular mesh.

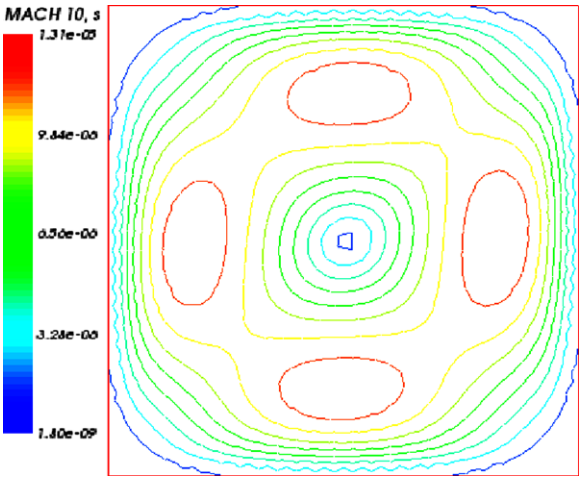

Fig. 27. Iso-Mach, *Low Mach Roe scheme*, 3D triangular mesh.

one of Figs. 21 and 23. Nevertheless, the proof in [36,37] does not take into account any boundary conditions. In the same way, the generalization of this proof to any type of triangular or tetrahedral mesh proposed in [35] does not take into account any boundary condition. Of course, boundary conditions (115) – especially the Dirichlet boundary condition applied to the temperature in the energy equation – are essential for the natural convection problem studied in this subsection, which perhaps explains why the numerical results are not good on Figs. 24 and 26 although the mesh is triangular. At the opposite, the *low Mach Roe scheme* remains accurate when the mesh is triangular (see Fig. 27) as well as when the mesh is Cartesian (see Fig. 20): these numerical results are a good surprise since all the theoretical results proposed in this paper are obtained for a periodic or unbounded domain Ω.

8. Conclusion

We have shown in the linear case that the convergence of a compressible Euler solution toward an incompressible Euler solution when the Mach number goes to zero is a direct consequence of the invariance of the well-prepared subspace

$$\mathcal{E} := \left\{ q := \begin{pmatrix} r \\ u \end{pmatrix} \in (L^2(\mathbb{T}^d))^{1+d} \text{ such that } \nabla r = 0 \text{ and } \nabla \cdot u = 0 \right\}$$

for the linear wave equation. Indeed, this invariance property avoids the creation of spurious (pressure and velocity) waves that are responsible for the loss of accuracy of Godunov type schemes at low Mach number in the non-linear case. Then, we have proposed to characterize a compressible colocated X scheme of Godunov type or not as being free of any spurious wave at low Mach number when the first-order modified equation associated to this X scheme applied to the linear wave equation preserves the well-prepared subspace $\mathcal{E}$ or a subspace $\mathcal{E}_{M\Delta x}$ that is an appropriate approximation of $\mathcal{E}$ when $M \ll 1$ and $M\Delta x \ll 1$.

Then, we have shown that a non-linear Godunov type scheme applied to the compressible Euler system on a 2D or 3D Cartesian mesh cannot be accurate at low Mach number because the first-order modified equation associated to the linearized Godunov scheme without convection does not preserve the well-prepared subspace $\mathcal{E}$ in 2D and in 3D and transfers energy of order 1 from the well-prepared subspace $\mathcal{E}$ toward the acoustic space $\mathcal{E}^{\perp}$ at an acoustic time scale. This loss of invariance is linked to the fact that the numerical diffusion operator in the velocity equation is non-isotropic. Nevertheless, the well-prepared subspace $\mathcal{E}$ remains invariant in 1D. This underlines that Godunov type schemes remain accurate at low Mach number in 1D. We have also underlined the importance of the space dimension with a formal analysis of the Roe scheme applied to the compressible Euler system. These statements emphasize that, firstly, the numerical diffusive terms on the mass and energy equations are not responsible for the loss of invariance in 2D and 3D at low Mach number. Secondly, this underlines that the inaccuracy can be explained (at least partly) with linear arguments and without any convection operator. Nevertheless, the non-linearities can amplify this inaccuracy through, for example, the non-linear convective operator and, thus, through shear waves. Moreover, boundary conditions can also contribute to the loss of invariance of the well-prepared subspace $\mathcal{E}$.

At the opposite, by deleting the numerical diffusion on the velocity equation of the first-order modified equation, we recover the invariance of the well-prepared subspace $\mathcal{E}$ in the linear case. As a consequence, we have proposed in the non-linear case and on any mesh the class of *low Mach X schemes* where X is a colocated scheme of Godunov type or not. A *low Mach X scheme* is simply built by centering the discretization of the pressure gradient, the rest of the scheme being unchanged. This simple modification is also justified by the fact that Roe–Turkel type schemes [4–9,13], the Roe scheme with pressure cor-

rection [10], the AUSM$^+$-up scheme [12] or the colocated schemes proposed in [16–18] are similar or identical at low Mach number to a *low Mach X scheme*. Moreover, by studying the L^2 stability in the linear case, we have noted that the time discretization of the *low Mach Godunov scheme* may be explicit as well as implicit (we recall that the Roe–Turkel scheme has to be implicited). At last, 2D numerical results show that the *low Mach Roe scheme* and the *low Mach VFRoe scheme* are stable and accurate at low Mach number on quadrangular and triangular meshes. Nevertheless, we emphasize that the *low Mach X schemes* are designed to give stable and accurate results *only* at low Mach number. But, it should be easy to adjust the deleting of the upwinding of the pressure gradient to be able to simulate flows whose Mach number is not everywhere or everytime close to zero. At last, let us underline that we have formally shown that there exists a link between the *low Mach Godunov scheme* and a colocated incompressible scheme stabilized with a Brezzi–Pitkäranta type stabilization [21].

Appendix A. Proof of Lemmas 2.4 and 6.1

Proof of Lemma 2.4: Since $\mathcal{A}$ is a linear isometry, there is conservation of the inner product that is to say $\forall (q_1,q_2)\in\mathbb{H}\times\mathbb{H}:\langle \mathcal{A}q_1,\mathcal{A}q_2\rangle=\langle q_1,q_2\rangle$. Thus, by taking $(q_1,q_2)\in\mathcal{E}\times\mathcal{E}^\perp$, we obtain that $\langle \mathcal{A}q_1,\mathcal{A}q_2\rangle=0$. Thus, we have $\mathcal{A}q_2\in(\mathcal{A}\mathcal{E})^\perp$ that is to say $\mathcal{A}q_2\in\mathcal{E}^\perp$ since $\mathcal{A}$ is supposed to be surjective. This proves that $\mathcal{A}\mathcal{E}^\perp\subset\mathcal{E}^\perp$. □

Proof of Lemma 6.1: Let us choose $q\in\mathbb{H}$. We have

$$(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q=(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu\mathbb{P}_\mu q+(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu(\mathbf{1}-\mathbb{P}_\mu)q$$

where $\mathbb{P}_\mu$ is the projection onto $\mathcal{E}_\mu$ along $\mathcal{E}^\perp$. Thus, we have

$$\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|\leqslant\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu\mathbb{P}_\mu q\|+\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu(\mathbf{1}-\mathbb{P}_\mu)q\|.$$

We know that $\mathcal{A}_\mu\mathbb{P}_\mu q\in\mathcal{E}_\mu$ since $\mathcal{A}_\mu\mathcal{E}_\mu\subset\mathcal{E}_\mu$. Thus, we have

$$\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu\mathbb{P}_\mu q\|\leqslant\mu|||\mathcal{A}_\mu\mathbb{P}_\mu q|||.$$

Moreover, $\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu(\mathbf{1}-\mathbb{P}_\mu)q\|\leqslant\|\mathcal{A}_\mu(\mathbf{1}-\mathbb{P}_\mu)q\|$ since $\mathbb{P}$ is an orthogonal projection. And by using the fact that $\mathcal{A}_\mu$ is bounded by λ for the norm $\|\cdot\|$, we obtain that $\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu(\mathbf{1}-\mathbb{P}_\mu)q\|\leqslant\lambda\|(\mathbf{1}-\mathbb{P}_\mu)q\|$. And since $\|(\mathbf{1}-\mathbb{P}_\mu)q\|\leqslant\|(\mathbb{P}-\mathbb{P}_\mu)q\|+\|(\mathbf{1}-\mathbb{P})q\|$, we can write that

$$\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|\leqslant\mu|||\mathcal{A}_\mu\mathbb{P}_\mu q|||+\lambda\|(\mathbb{P}-\mathbb{P}_\mu)q\|+\lambda\|(\mathbf{1}-\mathbb{P})q\|.$$

Moreover, we have $\mathbb{P}=\mathbb{P}\mathbb{P}_\mu$ since $\mathbb{P}$ and $\mathbb{P}_\mu$ are projections along the same subspace. Then, we have $(\mathbb{P}-\mathbb{P}_\mu)q=(\mathbb{P}-\mathbf{1})\mathbb{P}_\mu q$ that implies $\|(\mathbb{P}-\mathbb{P}_\mu)q\|\leqslant\mu|||\mathbb{P}_\mu q|||$ since $\mathbb{P}_\mu q\in\mathcal{E}_\mu$. Thus, we obtain that

$$\begin{aligned}\|(\mathbf{1}-\mathbb{P})\mathcal{A}_\mu q\|&\leqslant\mu(|||\mathcal{A}_\mu\mathbb{P}_\mu q|||+\lambda|||\mathbb{P}_\mu q|||)+\lambda\|(\mathbf{1}-\mathbb{P})q\|\leqslant\mu(|||\mathcal{A}_\mu|||+\lambda)|||\mathbb{P}_\mu q|||+\lambda\|(\mathbf{1}-\mathbb{P})q\|\\&\leqslant 2\lambda\mu|||\mathbb{P}_\mu q|||+\lambda\|(\mathbf{1}-\mathbb{P})q\|\quad\text{when}\quad\mu\in[0,\varepsilon[\end{aligned}$$

by using also the fact that $\mathcal{A}_\mu$ is bounded by λ for the second norm $|||\cdot|||$. This gives inequality (108). Moreover, we have

$$\mathbb{P}_\mu q=\mathbb{P}\mathbb{P}_\mu q+(\mathbf{1}-\mathbb{P})\mathbb{P}_\mu q=\mathbb{P}q+(\mathbf{1}-\mathbb{P})\mathbb{P}_\mu q\tag{116}$$

since $\mathbb{P}=\mathbb{P}\mathbb{P}_\mu$. Let us now suppose that there exists $\beta>0$ such that $\forall q\in\mathbb{H}:|||q|||\leqslant\beta\|q\|$. Then, inequality (107)(b) and relation (116) imply that $|||\mathbb{P}_\mu q|||\leqslant\beta\|\mathbb{P}q\|+\beta\mu|||\mathbb{P}_\mu q|||$ that is to say $|||\mathbb{P}_\mu q|||\leqslant\frac{\beta}{1-\beta\mu}\|\mathbb{P}q\|$ when $\mu\in[0,\min(\varepsilon,\beta^{-1})[$. We obtain inequality (109) by noting that $\|\mathbb{P}q\|\leqslant\|q\|$ since $\mathbb{P}$ is an orthogonal projection. □

Appendix B. Proof of Lemmas 5.1 and 5.2

Here, we define $a:=a_*/M$ to simplify the notations. A necessary condition for L^2 stability is that $|\lambda^\pm|\leqslant 1$ where $\lambda^\pm$ are the eigenvalues of the amplification matrix of each scheme in the Fourier space: this condition is the von Neumann necessary condition [27].

Proof of Lemma 5.1: In the Fourier space, the amplification matrix of explicit scheme (78) is given by

$$G_k^{exp}(\Delta t,\kappa)=\begin{pmatrix}1-\dfrac{2a\Delta t}{\Delta x}\kappa_r\sin^2\left(\dfrac{k\Delta x}{2}\right) & -\dfrac{a\Delta t}{\Delta x}\mathbf{i}\sin(k\Delta x)\\ -\dfrac{a\Delta t}{\Delta x}\mathbf{i}\sin(k\Delta x) & 1-\dfrac{2a\Delta t}{\Delta x}\kappa_u\sin^2\left(\dfrac{k\Delta x}{2}\right)\end{pmatrix}\tag{117}$$

($\mathbf{i}$ is the imaginary number such that $\mathbf{i}^2=-1$). We have

$$
\begin{aligned}
det(G_k^{exp}(\Delta t,\kappa)-\lambda I) &= (1-\lambda)^2-(1-\lambda)\frac{2a\Delta t}{\Delta x}(\kappa_r+\kappa_u)\sin^2\left(\frac{k\Delta x}{2}\right)+4\left(\frac{a\Delta t}{\Delta x}\right)^2\kappa_r\kappa_u\sin^4\left(\frac{k\Delta x}{2}\right)+\left(\frac{a\Delta t}{\Delta x}\right)^2\sin(k\Delta x)^2\\
&= (1-\lambda)^2-(1-\lambda)\frac{2a\Delta t}{\Delta x}(\kappa_r+\kappa_u)\sin^2\left(\frac{k\Delta x}{2}\right)+4\left(\frac{a\Delta t}{\Delta x}\right)^2\\
&\quad\times\sin^2\left(\frac{k\Delta x}{2}\right)\left[\kappa_r\kappa_u\sin^2\left(\frac{k\Delta x}{2}\right)+\cos^2\left(\frac{k\Delta x}{2}\right)\right]\\
&= (1-\lambda)^2-(1-\lambda)\frac{2a\Delta t}{\Delta x}(\kappa_r+\kappa_u)\sin^2\left(\frac{k\Delta x}{2}\right)+4\left(\frac{a\Delta t}{\Delta x}\right)^2\\
&\quad\times\sin^2\left(\frac{k\Delta x}{2}\right)\left[(\kappa_r\kappa_u-1)\sin^2\left(\frac{k\Delta x}{2}\right)+1\right]\\
&= \lambda^2+2\lambda\left[\frac{a\Delta t}{\Delta x}(\kappa_r+\kappa_u)\sin^2\left(\frac{k\Delta x}{2}\right)-1\right]+1+2\frac{a\Delta t}{\Delta x}\\
&\quad\times\sin^2\left(\frac{k\Delta x}{2}\right)\left\{2\frac{a\Delta t}{\Delta x}\left[(\kappa_r\kappa_u-1)\sin^2\left(\frac{k\Delta x}{2}\right)+1\right]-(\kappa_r+\kappa_u)\right\}
\end{aligned}
$$

that is to say

$$
det\left(G_k^{exp}(\Delta t,\kappa)-\lambda I\right)=\lambda^2+2\lambda[\mu(\kappa_r+\kappa_u)\theta^2-1]+1+2\mu\theta^2\{2\mu[(\kappa_r\kappa_u-1)\theta^2+1]-(\kappa_r+\kappa_u)\}
$$

with

$$
\begin{cases}
\mu=\dfrac{a\Delta t}{\Delta x},\\
\theta=|\sin\left(\dfrac{k\Delta x}{2}\right)|.
\end{cases}
$$

The discriminant Δ' of the polynomial $\lambda\mapsto det(G_k^{exp}(\Delta t,\kappa)-\lambda I)$ is given by

$$
\begin{aligned}
\Delta' &= [\mu(\kappa_r+\kappa_u)\theta^2-1]^2-1-2\mu\theta^2\{2\mu[(\kappa_r\kappa_u-1)\theta^2+1]-(\kappa_r+\kappa_u)\}\\
&= \mu^2(\kappa_r+\kappa_u)^2\theta^4+1-2\mu(\kappa_r+\kappa_u)\theta^2-1-4\mu^2\theta^2[(\kappa_r\kappa_u-1)\theta^2+1]+2\mu\theta^2(\kappa_r+\kappa_u)\\
&= \mu^2(\kappa_r+\kappa_u)^2\theta^4-4\mu^2\theta^2[(\kappa_r\kappa_u-1)\theta^2+1]=4\mu^2\theta^2\left[\frac{(\kappa_r+\kappa_u)^2}{4}\theta^2-(\kappa_r\kappa_u-1)\theta^2-1\right]\\
&= 4\mu^2\theta^2\left\{\left[\frac{(\kappa_r-\kappa_u)^2}{4}+1\right]\theta^2-1\right\}.
\end{aligned}
$$

Then, the eigenvalues λ^+ and λ^- are given by

$$
\lambda^{\pm}=-[\mu(\kappa_r+\kappa_u)\theta^2-1]\pm2\mu\theta\sqrt{\left[\frac{(\kappa_r-\kappa_u)^2}{4}+1\right]\theta^2-1}=1-\mu\theta^2\left[(\kappa_r+\kappa_u)\pm\sqrt{\frac{[(\kappa_r-\kappa_u)^2+4]\theta^2-4}{\theta^2}}\right]
$$

that is to say

$$
\lambda^{\pm}=1-\mu(\kappa_r+\kappa_u)\theta^2\left[1\pm\sqrt{\frac{[(\kappa_r-\kappa_u)^2+4]\theta^2-4}{(\kappa_r+\kappa_u)^2\theta^2}}\right]. \tag{118}
$$

- **First case:** $\kappa_r\in\mathbb{R}_*^+$ and $\kappa_u=\kappa_r$. Then

$$
\lambda^{\pm}=1-2\mu\kappa_r\theta^2\left[1\pm\frac{\mathbf{i}}{\kappa_r}\sqrt{\frac{1-\theta^2}{\theta^2}}\right].
$$

Thus

$$
|\lambda^{\pm}|^2=1+4\mu^2\theta^4\kappa_r^2-4\mu\theta^2\kappa_r+4\mu^2\theta^2(1-\theta^2)
$$

that is to say

$$|\lambda^{\pm}|^2 = 1 + 4\mu\theta^2\{\mu[1-\theta^2+\theta^2\kappa_r^2] - \kappa_r\}. \tag{119}$$

This proves that $|\lambda^{\pm}| < 1$ when $\mu \leqslant \frac{\kappa_r}{1-\theta^2+\theta^2\kappa_r^2}$. But

$$1 - \theta^2 + \theta^2\kappa_r^2 \leqslant \max(1, \kappa_r^2).$$

Thus, the stability condition may be written with $\mu \leqslant \frac{\kappa_r}{\max(1,\kappa_r^2)}$ that is to say with

$$\mu \leqslant \min(\kappa_r, \kappa_r^{-1}).$$

- **Second case:** $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = 0$. Then

$$\lambda^{\pm} = 1 - \mu\kappa_r\theta^2\left[1 \pm \sqrt{\frac{(\kappa_r^2+4)\theta^2 - 4}{\kappa_r^2\theta^2}}\right]. \tag{120}$$

Let us choose $\theta^2 \in \left[0, \frac{4}{\kappa_r^2+4}\right]$. Then

$$\lambda^{\pm} = 1 - \mu\kappa_r\theta^2\left[1 \pm \mathbf{i}\frac{\sqrt{4-(\kappa_r^2+4)\theta^2}}{\kappa_r\theta}\right].$$

Thus

$$|\lambda^{\pm}|^2 = (1-\mu\theta^2\kappa_r)^2 + \mu^2\theta^2[4-(\kappa_r^2+4)\theta^2] = 1 + \mu^2\theta^4\kappa_r^2 - 2\mu\theta^2\kappa_r + 4\mu^2\theta^2 - \mu^2\theta^4\kappa_r^2 - 4\mu^2\theta^4$$

that is to say

$$|\lambda^{\pm}|^2 = 1 + 2\mu\theta^2[2\mu(1-\theta^2) - \kappa_r]. \tag{121}$$

Thus, the stability condition is given by $\mu \leqslant \dfrac{\kappa_r}{2(1-\theta^2)}$ that is to say $\mu \leqslant \frac{\kappa_r}{2}$. Let us now choose $\theta^2 \in \left[\dfrac{4}{\kappa_r^2+4}, 1\right]$. Then $\lambda^{\pm} \in \mathbb{R}$. But, we have $\dfrac{(\kappa_r^2+4)\theta^2-4}{\kappa_r^2\theta^2} = \dfrac{\kappa_r^2\theta^2 + 4(\theta^2-1)}{\kappa_r^2\theta^2} \in [0,1]$ since $\theta^2 - 1 < 0$. Thus, $1 \pm \sqrt{\dfrac{(\kappa_r^2+4)\theta^2-4}{\kappa_r^2\theta^2}} > 0$ and the stability condition applied to (120) is given by

$$\mu \leqslant \frac{2}{\kappa_r\theta^2\left[1 \pm \sqrt{\dfrac{(\kappa_r^2+4)\theta^2-4}{\kappa_r^2\theta^2}}\right]}.$$

But, the function $\theta^2 \mapsto \theta^2\left[1 \pm \sqrt{\dfrac{(\kappa_r^2+4)\theta^2-4}{\kappa_r^2\theta^2}}\right]$ is positive and bounded by 2 on $\left[\dfrac{4}{\kappa_r^2+4}, 1\right]$. Thus, the stability condition is given by

$$\mu \leqslant \frac{1}{\kappa_r}.$$

To summarize, we have obtained the stability condition

$$\mu \leqslant \min\left(\frac{\kappa_r}{2}, \frac{1}{\kappa_r}\right). \qquad \square$$

Proof of Lemma 5.2: The eigenvalues $\lambda^{\pm}$ of the amplification matrix $G_k^{imp}(\Delta t, \kappa)$ associated to implicit scheme (81) are such that $1/\lambda^{\pm}$ are the eigenvalues of

$$[G_k^{imp}(\Delta t, \kappa)]^{-1} = \begin{pmatrix} 1 + \dfrac{2a\Delta t}{\Delta x}\kappa_r \sin^2\left(\dfrac{k\Delta x}{2}\right) & \dfrac{a\Delta t}{\Delta x}\mathbf{i}\sin(k\Delta x) \\ \dfrac{a\Delta t}{\Delta x}\mathbf{i}\sin(k\Delta x) & 1 + \dfrac{2a\Delta t}{\Delta x}\kappa_u \sin^2\left(\dfrac{k\Delta x}{2}\right) \end{pmatrix}. \tag{122}$$

The eigenvalues $1/\lambda^{\pm}$ of $[G_k^{imp}(\Delta t, \kappa)]^{-1}$ are deduced from those of $G_k^{exp}(\Delta t, \kappa)$ by replacing μ with $-\mu$ in formula (118). Thus, we have

$$\lambda^{\pm} = \frac{1}{1 + \mu(\kappa_r+\kappa_u)\theta^2\left[1 \pm \sqrt{\dfrac{[(\kappa_r-\kappa_u)^2+4]\theta^2-4}{(\kappa_r+\kappa_u)^2\theta^2}}\right]}.$$

- **First case:** $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = \kappa_r$. We deduce from (119) that

$$|\lambda^\pm|^2 = \frac{1}{1 - 4\mu\theta^2\{-\mu[1-\theta^2+\theta^2\kappa_r^2] - \kappa_r\}} = \frac{1}{1 + 4\mu\theta^2\{\mu[1-\theta^2+\theta^2\kappa_r^2] + \kappa_r\}}.$$

Thus, we have $|\lambda^\pm| \leqslant 1$ for any $\mu \geqslant 0$.

- **Second case:** $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = 0$. Let us choose $\theta^2 \in [0, \frac{4}{\kappa_r^2+4}]$. Then, we deduce from (121) that

$$|\lambda^\pm|^2 = \frac{1}{1 - 2\mu\theta^2[-2\mu(1-\theta^2) - \kappa_r]} = \frac{1}{1 + 2\mu\theta^2[2\mu(1-\theta^2) + \kappa_r]}.$$

Thus, we have $|\lambda^\pm| \leqslant 1$ for any $\mu \geqslant 0$. Let us now choose $\theta^2 \in \left[\frac{4}{\kappa_r^2+4}, 1\right]$. Then, $\lambda^\pm \in \mathbb{R}$ since

$$\lambda^\pm = \frac{1}{1 + \mu\kappa_r\theta^2\left[1 \pm \sqrt{\dfrac{(\kappa_r^2+4)\theta^2 - 4}{\kappa_r^2\theta^2}}\right]}.$$

But, we know that the function $\theta^2 \mapsto \theta^2\left[1 \pm \sqrt{\dfrac{(\kappa_r^2+4)\theta^2-4}{\kappa_r^2\theta^2}}\right]$ is positive on $\left[\dfrac{4}{\kappa_r^2+4}, 1\right]$ (see proof of Lemma 5.1). Thus, we have also $0 < \lambda^\pm \leqslant 1$ for any $\mu \geqslant 0$. □

Remark on the von Neumann condition when $\kappa_r \in \mathbb{R}_*^+$ and $\kappa_u = \kappa_r$: The von Neumann condition is a necessary condition to obtain the L^2 stability: see condition S of the Kreiss matrix theorem (see [27], p. 74). Nevertheless, we can prove that $PG_k^{exp}P^{-1}$ is diagonale where G_k^{exp} is given by (117) and where

$$P = P^{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

when $sin(k\Delta x) \neq 0$. Moreover, when $sin(k\Delta x) = 0$, we have $\lambda^+ = \lambda^- := \lambda$ and $G_k = \lambda I$ (where I is the identity matrix). This proves that condition S of the Kreiss matrix theorem is satisfied for the amplification matrix G_k^{exp}. We obtain the same result for the amplification matrix G_k^{imp} defined with (122) since G_k^{imp} is also diagonalizable with P.

Appendix C. Proof of Proposition 5.1

Point 1: The Roe–Turkel scheme applied to the linear wave equation

$$\partial_t q + \frac{L}{M} q = 0 \tag{123}$$

is given by

$$\frac{d}{dt} q_{ij} + \frac{\widehat{F}_{1,i+1/2,j} - \widehat{F}_{1,i-1/2,j}}{\Delta x_1} + \frac{\widehat{F}_{2,ij+1/2} - \widehat{F}_{2,ij-1/2}}{\Delta x_2} = 0 \tag{124}$$

where

$$\begin{cases} \widehat{F}_{1,i+1/2,j} = \widehat{A}_1 \dfrac{q_{ij} + q_{i+1,j}}{2} - \dfrac{1}{2}\widehat{P}^{-1}|\widehat{P}\widehat{A}_1|(q_{i+1,j} - q_{ij}), \\ \widehat{F}_{2,ij+1/2} = \widehat{A}_2 \dfrac{q_{ij} + q_{ij+1}}{2} - \dfrac{1}{2}\widehat{P}^{-1}|\widehat{P}\widehat{A}_2|(q_{ij+1} - q_{ij}) \end{cases} \tag{125}$$

with

$$\widehat{A}_1 = \frac{a_*}{M}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \widehat{A}_2 = \frac{a_*}{M}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

In (125), $\widehat{P}$ is the Turkel preconditioning matrix [14,15] defined with

$$\widehat{P} = \begin{pmatrix} M^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The spectral decomposition of $\widehat{P}\widehat{A}_1 = \frac{a_*}{M}\begin{pmatrix} 0 & M^2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ is given by

$$\begin{cases} \lambda_1 = -a_* : & r_1 = \sqrt{\frac{M}{2}} \begin{pmatrix} 1 \\ -\frac{1}{M} \\ 0 \end{pmatrix}, \quad l_1 = \sqrt{\frac{M}{2}} \begin{pmatrix} \frac{1}{M} \\ -1 \\ 0 \end{pmatrix} \\ \lambda_2 = a_* : & r_2 = \sqrt{\frac{M}{2}} \begin{pmatrix} 1 \\ \frac{1}{M} \\ 0 \end{pmatrix}, \quad l_2 = \sqrt{\frac{M}{2}} \begin{pmatrix} \frac{1}{M} \\ 1 \\ 0 \end{pmatrix} \\ \lambda_3 = 0 : & r_3 = l_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{cases}$$

Thus

$$|\widehat{P}\widehat{A}_1| = a_* \cdot (r_1 \otimes l_1 + r_2 \otimes l_2) = \frac{a_* M}{2} \left[\begin{pmatrix} \frac{1}{M} & -1 & 0 \\ -\frac{1}{M^2} & \frac{1}{M} & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{M} & 1 & 0 \\ \frac{1}{M^2} & \frac{1}{M} & 0 \\ 0 & 0 & 0 \end{pmatrix} \right] = a_* \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Finally, we obtain

$$\widehat{P}^{-1}|\widehat{P}\widehat{A}_1| = \frac{a_*}{M} \begin{pmatrix} \frac{1}{M} & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{126}$$

A similar calculation gives

$$\widehat{P}^{-1}|\widehat{P}\widehat{A}_2| = \frac{a_*}{M} \begin{pmatrix} \frac{1}{M} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M \end{pmatrix}. \tag{127}$$

Thus, scheme (124) and (125) can be rewritten with

$$\begin{aligned} &\frac{d}{dt} q_{ij} + \widehat{A}_1 \frac{q_{i+1j} - q_{i-1j}}{2\Delta x_1} + \widehat{A}_2 \frac{q_{ij+1} - q_{ij-1}}{2\Delta x_2} \\ &\quad = \cdots = \frac{a_*}{2M} \cdot \left[\begin{pmatrix} \frac{1}{M} & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{q_{i+1j} - 2q_{ij} + q_{i-1j}}{\Delta x_1} + \begin{pmatrix} \frac{1}{M} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M \end{pmatrix} \frac{q_{ij+1} - 2q_{ij} + q_{ij-1}}{\Delta x_2} \right]. \end{aligned} \tag{128}$$

We conclude by noting that the first-order modified equation associated to scheme (128) is given by (61) where $B_\kappa q$ is defined with (89).

Point 2: The Roe scheme with pressure correction applied to linear wave equation (123) is defined with (124) and

$$\begin{cases} \widehat{F}_{1,i+1/2j} = \widehat{A}_1 \frac{q_{ij} + q_{i+1j}}{2} - \frac{1}{2} |\widehat{A}_1| (q_{i+1j} - q_{ij}) + \frac{a_*}{2M} \begin{pmatrix} 0 \\ u_{1,i+1j} - u_{1,ij} \\ 0 \end{pmatrix}, \\ \widehat{F}_{2,ij+1/2} = \widehat{A}_2 \frac{q_{ij} + q_{ij+1}}{2} - \frac{1}{2} |\widehat{A}_2| (q_{ij+1} - q_{ij}) + \frac{a_*}{2M} \begin{pmatrix} 0 \\ 0 \\ u_{2,ij+1} - u_{2,ij} \end{pmatrix}. \end{cases} \tag{129}$$

The corrected fluxes (129) are deduced from (42) restricted to the case of linear wave equation (123). By taking $\widehat{P} = I$ in (126) and (127), we obtain

$$\frac{1}{2} |\widehat{A}_1| (q_{i+1j} - q_{ij}) = \frac{a_*}{2M} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} (q_{i+1j} - q_{ij})$$

and

$$\frac{1}{2} |\widehat{A}_2| (q_{ij+1} - q_{ij}) = \frac{a_*}{2M} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} (q_{ij+1} - q_{ij}).$$

Moreover, we have

$$\frac{a_*}{2M}\begin{pmatrix}0\\ u_{1,i+1,j}-u_{1,i,j}\\ 0\end{pmatrix}=\frac{a_*}{2M}\begin{pmatrix}0&0&0\\0&1&0\\0&0&0\end{pmatrix}(q_{i+1,j}-q_{i,j})$$

and

$$\frac{a_*}{2M}\begin{pmatrix}0\\ 0\\ u_{2,i,j+1}-u_{2,i,j}\end{pmatrix}=\frac{a_*}{2M}\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}(q_{i,j+1}-q_{i,j}).$$

Thus, scheme (124) (129) can be rewritten with

$$\frac{d}{dt}q_{i,j}+\widehat{A}_1\frac{q_{i+1,j}-q_{i-1,j}}{2\Delta x_1}+\widehat{A}_2\frac{q_{i,j+1}-q_{i,j-1}}{2\Delta x_2}=\cdots=\frac{a_*}{2M}\cdot\left[\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}\frac{q_{i+1,j}-2q_{i,j}+q_{i-1,j}}{\Delta x_1}+\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}\frac{q_{i,j+1}-2q_{i,j}+q_{i,j-1}}{\Delta x_2}\right]. \tag{130}$$

We conclude by noting that the first-order modified equation associated to scheme (130) is given by (61) where $B_\kappa q$ is defined with (90).

References

[1] P.L. Roe, Approximate Riemann solvers, parameter vectors and difference schemes, J. Comp. Phys. 43 (1981) 357–372.
[2] J.-M. Massella, I. Faille, T. Gallouët, On an approximate Godunov scheme, Int. J. Comput. Fluid. Dyn. 12 (1999) 133–149.
[3] T. Buffard, T. Gallouët, J.-M. Hérard, A sequel to a rough Godunov scheme: application to real gases, Comput. Fluids 29 (2000) 813–847.
[4] F. Beux, M.V. Salvetti, E. Sinibaldi, A preconditioned implicit Roe's scheme for barotropic flows: towards simulation of cavitation phenomena, Technical Report 4891, INRIA, 2003.
[5] S. Clerc, Numerical simulation of the homogeneous equilibrium model for two-phase flow, J. Comp. Phys. 161 (2000) 354–375.
[6] F. Dauvergne, J.-M. Ghidaglia, F. Pascal, J.-M. Rovarch, Renormalization of the numerical diffusion for an upwind finite volume method. Application to the simulation of Kelvin–Helmholtz instability, in: R. Eymard, J.-M. Hérard (Eds.), Proceedings of the Fifth International Symposium on Finite Volumes for Complex Applications, Wiley, 2008, pp. 321–328.
[7] H. Guillard, C. Viozat, On the behavior of upwind schemes in the low Mach number limit, Comput. Fluids 28 (1999) 63–86.
[8] H. Guillard, A. Murrone, On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes, Comput. Fluids 33 (2004) 655–675.
[9] H. Guillard, A. Murrone, Behavior of upwind scheme in the low Mach number limit: III. Preconditioned dissipation for a five equation two phase model, Comput. Fluids 37 (10) (2008) 1209–1224.
[10] H. Paillère, C. Viozat, A. Kumbaro, I. Toumi, Comparison of low Mach number models for natural convection problems, Heat and Mass Transfer 36 (2000) 567–573.
[11] G. Volpe, Performance of compressible flow codes at low Mach number, AIAA J. 31 (1) (1993) 49–56.
[12] M.-S. Liou, A sequel to AUSM: AUSM$^+$, Part II: AUSM$^+$-up for all speeds, J. Comp. Phys. 214 (1) (2006) 137–170.
[13] M. Sabanca, G. Brenner, N. Alemdaroglu, Improvements to compressible Euler methods for low-Mach number flows, Int. J. Numer. Meth. Fluids 34 (2000) 167–185.
[14] E. Turkel, Preconditioned methods for solving the incompressible and low speed compressible equations, J. Comp. Phys. 72 (1987) 277–298.
[15] E. Turkel, Review of preconditioning methods for fluid dynamics, Appl. Numer. Math. 12 (1993) 257–284.
[16] I. Mary, P. Sagaut, Large Eddy simulation of flow around an airfoil near stall, AIAA J. 40 (6) (2002) 1139–1145.
[17] X.-S. Li, C.-W. Gu, An all-speed Roe-type scheme and its asymptotic analysis of low Mach number behaviour, J. Comp. Phys. 227 (2008) 5144–5159.
[18] X.-S. Li, C.-W. Gu, J.-Z. Xu, Development of Roe-type scheme for all-speed flows based on preconditioning method, Comput. Fluids 38 (2009) 810–817.
[19] S. Schochet, Fast singular limits of hyperbolic PDEs, J. Differ. Eqs. 114 (1994) 476–512.
[20] B. Perthame, Boltzmann type schemes for the gas dynamics and the entropy property, SIAM J. Num. Anal. 27 (1990) 1405–1421.
[21] R. Eymard, R. Herbin, J.-C. Latché, On a stabilized colocated finite volume scheme for the Stokes problem, Math. Model. Numer. Anal. 40 (3) (2006) 501–527.
[22] A. Majda, in: Springer-Verlag (Ed.), Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables, Applied Mathematical Sciences Series, Wiley, New York, 1984, pp. 30–80.
[23] F. Boyer, P. Fabrie, in: Éléments d'analyse pour l'étude de quelques modèles d'écoulements de fluides visqueux incompressibles, Mathématiques & Applications, Springer, French, 2006 (Chapter 2).
[24] M.L. Minion, A projection method for locally refined grids, J. Comp. Phys. 127 (1996) 158–178.
[25] E. Godlewski, P.-A. Raviart, in: J. Mardsen, L. Sirovich, F. John (Eds.), Numerical Approximation of Hyperbolic Systems of Conservation Laws, vol. 118, Springer-Verlag, New York, 1996, pp. 215–220.
[26] B.J.R. Thornber, D. Drikakis, Numerical dissipation of upwind schemes in low Mach flow, Int. J. Numer. Meth. Fluids 56 (2008) 1535–1541.
[27] R.D. Richtmyer, K.W. Morton, in: Krieger (Ed.), Difference Methods for Initial-Value Problems, second ed., 1994.
[28] S. Dellacherie, Checkerboard modes and wave equation, in: Proceedings of the 18th Conference on Scientific Computing, Podbanske, Slovakia, 2009, pp. 71–80.
[29] A. Beccantini, Colella–Glaz splitting scheme for thermally perfect gases, in: E.F. Toro (Ed.), Godunov Methods: Theory and Application, Kluwer Academic/Plenum Publishers, New York, 2001, pp. 89–95.
[30] P. Colella, H. Glaz, Efficient solution algorithms for the Riemann problem for real gases, J. Comp. Phys. 59 (1985) 264–289.
[31] P. Birken, A. Meister, Stability of preconditioned finite volume schemes at low Mach numbers, BIT Numer. Math. 45 (3) (2005) 463–480.
[32] M.-S. Liou, A sequel to AUSM: AUSM$^+$, J. Comp. Phys. 129 (2) (1996) 364–382.
[33] M.-S. Liou, C.J. Steffen, A new flux splitting scheme, J. Comp. Phys. 107 (1) (1993) 23–39.
[34] F. Brezzi, J. Pitkäranta, On the Stabilization of Finite Element Approximation of the Stokes Equation, in: W. Hackbusch (Ed.), Efficient Solutions of Elliptic Systems, Notes on Numerical Fluid Mechanics, vol. 10, 1984, pp.11–19.
[35] S. Dellacherie, P. Omnes, F. Rieper, The influence of cell geometry on the Godunov scheme applied to the linear wave equation, submitted for publication.
[36] F. Rieper, Influence of cell geometry on the behaviour of the first-order roe scheme in the low Mach number regime, in: R. Eymard, J.-M. Hérard (Eds.), Finite Volumes for Complex Applications V, Wiley, 2008, pp. 625–632.

[37] F. Rieper, G. Bader, The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime, J. Comp. Phys. 228 (8) (2009) 2918–2933.
[38] D. Vidović, A. Segal, P. Wesseling, A superlinearly convergent Mach-uniform finite volume method for the Euler equations on staggered unstructured grids, J. Comp. Phys. 217 (2006) 277–294.
[39] G.D.V. Davis, Natural convection of air in a square cavity: a bench mark numerical solution, Int. J. Numer. Meth. Fluids 3 (1983) 249–264.
[40] P.L.Q. et al., Modeling of natural convection flows with large temperature differences: a benchmark problem for low Mach number solvers. Part 1. Reference Solutions, Math. Model. Numer. Anal. 39 (3) (2005a) 609–616.
[41] H.P. et al., Modeling of natural convection flows with large temperature differences: a benchmark problem for low Mach number solvers. Part 2. Contributions to the June 2004 Conference, Math. Model. Numer. Anal. 39 (3) (2005b) 617–621.