Flux Splitting: A Notion on Stability

Sebastian Noelle and Jochen Schütz

Bericht Nr. 382 **Januar 2014**

Key words: IMEX finite volume, asymptotic preserving, flux splitting, modified equation, stability analysis

AMS Subject Classifications: 36L65, 76M45, 65M08

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Templergraben 55, D–52056 Aachen (Germany)

S. Noelle and J. Schütz
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University,
Templergraben 55, 52062 Aachen
Tel: +49 241 80 97677
E–mail: {noelle,schuetz}@igpm.rwth-aachen.de

Journal of Scientific Computing manuscript No.
(will be inserted by the editor)

Flux Splitting: A notion on stability

Sebastian Noelle · Jochen Schütz

Received: date / Accepted: date

Abstract In the context of low Mach number flows, successful methods are *Asymptotic Preserving* and *IMEX* schemes. Both schemes for hyperbolic equations rely on a splitting of the convective flux into stiff and nonstiff parts. This choice is not arbitrary and has an influence on both the stability and the accuracy of the resulting methods. In this work, we consider a first-order IMEX scheme based on different splittings. Using the modified equation approach, we can show that a new class of splittings based on characteristic decomposition, also introduced in this work, gives rise to a stable method with a time step independent of the small quantity ε, whereas a splitting taken from literature can be identified to be stable only for small values of the CFL number.

1 Introduction, underlying equations and flux splitting

In recent years there has been a renewed interest in the computation of singularly perturbed differential equations. These equations arise, e.g., in the simulation of low-speed fluid flows. Here we are interested in computing waves with vastly different speeds. The goal is to resolve slow waves accurately and efficiently with a large time step, while approximating the fast waves in a stable way, using the same time step.

A class of algorithms that has found particular attention are the *Asymptotic Preserving* schemes introduced by Jin [9] building on work with Pareschi and Toscani [11]. For an excellent review article, consult [10]; we refer to [7,4,6,1, 14,16] for various applications of this method in different contexts.

All the algorithms used by the authors cited above rely on identifying *stiff* and *nonstiff* parts of the underlying equation. This point is generally

S. Noelle and J. Schütz
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University
Templergraben 55, 52062 Aachen
Tel.: +49 241 80 97677
E-mail: {noelle,schuetz}@igpm.rwth-aachen.de

considered crucial as a well-chosen splitting guarantees a good behavior of the algorithm. A splitting is usually obtained by physical reasoning, see e.g. the fundamental work by Klein [12].

Having obtained a splitting into stiff and nonstiff parts, the nonstiff part is then treated explicitly, and the stiff one implicitly. This procedure naturally leads to so-called IMEX schemes as introduced in [2], see also [3,15] for an interesting discussion on the quality of these schemes in the asymptotic limit. In [8] the authors show that, if both the explicit and the implicit part (considered *separately*) are stable, then so is the full algorithm. However, their procedure does not yield quantitative information on the CFL numbers needed for stability.

In this work, we investigate stability of the lowest-order IMEX scheme for *linear* equations using the (heuristic) idea of the *modified equation* approach, see [18]. We derive the modified parabolic system of equations of second order and investigate under what conditions its solutions are damped. For simple problems, we can investigate this analytically, for more involved problems, we use numerical examples. The idea of the modified equation is closely related to L^2-stability, often also called *linear* stability, introduced in [13] as the famous *von-Neumann* analysis. Strang [17] showed that, under some assumptions, it is enough to consider only linearized problems, so the approach used in this work is actually more general than it seems on first sight.

In this publication, we consider the linear system of conservation laws

$$u_t + Au_x = 0 \qquad \forall (x,t) \in \Omega \times (0,T) \tag{1}$$

$$u(x,0) = u_0(x) \quad \forall x \in \Omega \tag{2}$$

with a constant matrix $A \in \mathbb{R}^{d\times d}$ that has d distinct eigenvalues and a full set of corresponding eigenvectors. For simplicity, we consider a periodic setting with a period of one. Therefore, without loss of generality, we set $\Omega := [0,1]$. Furthermore, we assume that the matrix A is a function of a parameter $\varepsilon \in (0,1]$ such that with $\varepsilon \to 0$, some eigenvalues of A diverge towards infinity. We assume that u_0 is a smooth periodic function on $\mathbb{R}$ with unit periodicity, so that one can expect u to be smooth.

The motivation to consider (1) stems from considering *linearized* versions of classical systems of conservation laws

$$v_t + f(v)_x = 0, \tag{3}$$

e.g., the (non-dimensionalized) Euler equations at low Mach number ε, with $v = (\rho, \rho u, E)^T$ and

$$f(\rho, \rho u, E) := \left(\rho u, \rho u^2 + \frac{p}{\varepsilon^2}, u(E+p)\right)^T. \tag{4}$$

Its linearization around a state $(\rho, \rho u, E)$ yields a matrix A with eigenvalues

$$\lambda = u, u \pm \frac{c}{\varepsilon}. \tag{5}$$

Note that two eigenvalues tend to infinity as $\varepsilon \to 0$.

To highlight the difficulties posed by eigenvalues of multiple scales we discuss a standard, explicit finite volume scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n - \widehat{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} = 0 \tag{6}$$

with consistent numerical flux $\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n$. From the stability conditions by Courant, Friedrichs and Lewy [5], it is known that explicit schemes are only stable under a CFL condition, which is typically given by

$$\nu_{\max} := \lambda_{\max} \frac{\Delta t}{\Delta x} = \frac{(u + \frac{c}{\varepsilon})\Delta t}{\Delta x} < 1. \tag{7}$$

In the limit as $\varepsilon \to 0$, one mainly wants to resolve the advective wave traveling with speed u. Given the restrictive CFL condition (7), this would imply that one needs $\mathcal{O}(\varepsilon^{-1})$ steps to advect a signal across a single grid cell. For small ε, this is prohibitively inefficient, and for many schemes also prohibitively dissipative. However, using

$$\widehat{\nu} := u \frac{\Delta t}{\Delta x} < 1 \tag{8}$$

as *advective* CFL condition would result in an unstable scheme.

One potential remedy is to use fully implicit or mixed implicit / explicit (IMEX) methods. The latter class of methods requires a splitting of A into components with 'slow' and 'fast' waves. More specifically, one seeks matrices $\widehat{A}$ and $\widetilde{A}$, such that

$$A = \widehat{A} + \widetilde{A}, \tag{9}$$

with the following conditions posed on $\widehat{A}$ and $\widetilde{A}$:

Definition 1 The splitting (9) is called *admissible*, if

- both $\widehat{A}$ and $\widetilde{A}$ induce a hyperbolic system, i.e., they have real eigenvalues and a complete set of eigenvectors.
- the eigenvalues of $\widehat{A}$ are bounded independently of ε.

In this work, we give a recipe for identifying efficient and stable classes of flux splittings. We use the well-known modified equation analysis as a tool for (heuristically) investigating L^2-stability.

The paper is outlined as follows: In Section 2, we introduce so-called characteristic splittings, which will be basic ingredients for a stable scheme. In Section 3, we introduce the lowest-order IMEX scheme, while in Section 4, we investigate this scheme for stability using the modified equation approach. In Section 5, we show that the characteristic splittings are stable in the sense as explained in Section 4. Section 6 shows an example of an unstable scheme. Finally, Section 7 offers conclusions and future work.

2 Characteristic splitting

In this section, we introduce a new class of splittings that, with our analysis to be presented, turns out to be uniformly stable in ε *without* any additional stabilization terms. The splitting relies on a characteristic decomposition of the matrix A, i.e., A can be decomposed into

$$A = Q\Lambda Q^{-1} \tag{10}$$

for an invertible Q and $\Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_d)$. The idea of the characteristic splitting is to split the matrix Λ into stiff and nonstiff parts as

$$\Lambda = \widehat{\Lambda} + \widetilde{\Lambda}, \tag{11}$$

where $\widehat{\Lambda}$ and $\widetilde{\Lambda}$ are diagonal matrices and define an admissible splitting of Λ in the sense of Definition 1. Consequently, the entries of $\widehat{\Lambda}$ can be bounded independently of ε. Subsequently, the splitting is defined as

$$\widehat{A} = Q\widehat{\Lambda}Q^{-1} \quad \text{and} \quad \widetilde{A} = Q\widetilde{\Lambda}Q^{-1}. \tag{12}$$

Obviously, the splitting is admissible in the sense of Definition 1. In the sequel, we give two simple examples.

2.1 Prototype Matrix

We consider a prototype matrix A given by

$$A = \begin{pmatrix} a & 1 & 0 \\ \frac{1}{\varepsilon^2} & a & \frac{1}{\varepsilon^2} \\ 0 & 1 & a \end{pmatrix}. \tag{13}$$

Its eigenvalues are

$$\lambda = a, a \pm \frac{\sqrt{2}}{\varepsilon}, \tag{14}$$

and for simplicity, we consider a to be positive, i.e., $\lambda_{\max} := a + \frac{\sqrt{2}}{\varepsilon}$ is the largest eigenvalue.

In order to be fully explicit for $\varepsilon = 1$, we use a characteristic splitting with

$$\widehat{\Lambda} := \mathrm{diag}\left(a - \sqrt{2}, a, a + \sqrt{2}\right), \tag{15}$$

$$\widetilde{\Lambda} := \mathrm{diag}\left(-\frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}, 0, \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}\right). \tag{16}$$

Consequently, we can derive matrices $\widehat{A}$ and $\widetilde{A}$ via (12) as

$$\widehat{A} = \begin{pmatrix} a & \varepsilon & 0 \\ \frac{1}{\varepsilon} & a & \frac{1}{\varepsilon} \\ 0 & \varepsilon & a \end{pmatrix}, \quad \widetilde{A} = \begin{pmatrix} 0 & 1-\varepsilon & 0 \\ \frac{1-\varepsilon}{\varepsilon^2} & 0 & \frac{1-\varepsilon}{\varepsilon^2} \\ 0 & 1-\varepsilon & 0 \end{pmatrix}. \tag{17}$$

2.2 Euler equations

Here, we give a straightforward splitting of the linearization of equation (3) with Euler fluxes (4). As the matrix $A := f'(\rho, \rho u, E)$ is hyperbolic for any choice of $(\rho, \rho u, E)$, it can also be diagonalized. Its eigenvalues are given in (5) as

$$\lambda = u, u \pm \frac{c}{\varepsilon} \tag{18}$$

for $c := \sqrt{\frac{\gamma p}{\rho}}$. A straightforward splitting is given by using the diagonal matrices

$$\widehat{\Lambda} := \operatorname{diag}(u - c, u, u + c), \tag{19}$$

$$\widetilde{\Lambda} := \operatorname{diag}\left(-\frac{c(1-\varepsilon)}{\varepsilon}, 0, \frac{c(1-\varepsilon)}{\varepsilon}\right). \tag{20}$$

Again, the stiff part vanishes for $\varepsilon = 1$; $\widehat{A}$ and $\widetilde{A}$ are defined via (12).

3 IMEX discretization

Based on a splitting as given in (9), we introduce a straightforward first-order IMEX discretization based on nonstiff and stiff numerical fluxes $\widehat{\mathcal{H}}$ and $\widetilde{\mathcal{H}}$. We assume that the temporal domain is subdivided as

$$0 = t^0 < t^1 < \ldots < t^N = T \tag{21}$$

with constant spacing $\Delta t := t^{n+1} - t^n$; and that we have a subdivision

$$\Omega := \bigcup_{j=0}^{J} [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \tag{22}$$

also with constant spacing $\Delta x := x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ and cell midpoints x_j. As is customary, we denote an approximation to $u(x_j, t^n)$ by u_j^n. Furthermore, the vector $(u_0^n, \ldots u_J^n)$ is denoted by u^n. Now we can introduce a (classical) first-order IMEX scheme:

Definition 2 A sequence $(u^0, \ldots, u^N)$ is a solution to an IMEX discretization, given that

$$\mathcal{I}_j^n(u^n, u^{n+1}) := \frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n - \widehat{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} + \frac{\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} - \widetilde{\mathcal{H}}_{j-\frac{1}{2}}^{n+1}}{\Delta x} = 0 \quad \forall j, n. \tag{23}$$

Here, nonstiff and stiff numerical fluxes are defined by

$$\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n := \frac{1}{2}\widehat{A}\left(u_{j+1}^n + u_j^n\right) \quad - \frac{\widehat{\alpha}}{2}\left(u_{j+1}^n - u_j^n\right) \tag{24}$$

$$\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} := \frac{1}{2}\widetilde{A}\left(u_{j+1}^{n+1} + u_j^{n+1}\right) - \frac{\widetilde{\alpha}}{2}\left(u_{j+1}^{n+1} - u_j^{n+1}\right), \tag{25}$$

with (positive) numerical viscosities $\widehat{\alpha}$ and $\widetilde{\alpha}$.

For fixed ε, consistency analysis of the scheme is well-known. However, we have to consider both ε and Δt as small parameters. The crucial point is that we restrict our analysis to cases where the magnitude of u and its derivatives are independent of ε. (Especially, no derivative behaves as $O(\varepsilon^{-1})$ or worse.) This assumption is reasonable, as only these solutions allow for an asymptotic limit as $\varepsilon \to 0$.

Lemma 1 *Let U^n denote the vector $(u(x_0,t^n),\dots,u(x_J,t^n))$ with u solution to* (1) *whose derivatives can be bounded* independently *of ε; and similarly for U^{n+1}. Furthermore, let $O(\Delta x) = O(\Delta t)$. Then, the local truncation error is of order Δx, i.e., there holds*

$$\mathcal{I}_j^n(U^n,U^{n+1}) = O(\Delta x). \tag{26}$$

Proof Apply a Taylor expansion to (23) *and note that all the derivatives of u can be bounded independently of ε.*

The focus of this paper is on uniform stability as $\varepsilon \to 0$, where the fast wave speeds tend to infinity. As outlined in the introduction, the goal is to overcome the inefficiency of a fully explicit scheme due to condition (7), or the instability due to condition (8). In the following (cf. Lemma 3), we will derive upper bounds on the nonstiff CFL number that assure stability (in a sense to be made more precise) of IMEX scheme (23) for a characteristic splitting.

4 Modified equation analysis

In this section, we derive the modified equation [18] corresponding to (23). As we consider a periodic setting, we can solve the resulting parabolic system explicitly using Fourier series. Using Plancherel's theorem, we investigate the stability of the modified equation. This yields a practical criterion for the stability of the IMEX scheme.

We start by deriving the modified equation corresponding to (23).

Theorem 1 *Let w be a smooth solution of*

$$w_t + Aw_x = \frac{\Delta t}{2}\left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t}\,\mathrm{Id} - (\widehat{A}-\widetilde{A})A\right) w_{xx}. \tag{27}$$

Furthermore, we consider vector $W^n := (w(x_0,t^n),\dots,w(x_J,t^n))$. Then, for fixed ε and $O(\Delta x) = O(\Delta t)$, the IMEX scheme (23) *is a second order accurate discretization of* (27)*, i.e.*

$$\mathcal{I}_j^n(W^n,W^{n+1}) = O(\Delta x^2). \tag{28}$$

Proof It is well-known that the modified equation for a first-order discretization is a parabolic equation, i.e., we expect w to fulfill

$$w_t + Aw_x = Bw_{xx} \tag{29}$$

for a (yet unknown) viscosity matrix B that is in class $O(\Delta x)$. Applying the Cauchy-Kovalewskaya expansion to (29) *gives*

$$w_t = -Aw_x + Bw_{xx} \tag{30}$$

$$w_{tt} = -A(w_t)_x + B(w_t)_{xx} \overset{(30)}{=} A^2 w_{xx} + O(\Delta t). \tag{31}$$

To simplify the presentation, we slightly abuse our notation, and write w_j^n for $w(x_j, t^n)$. Using (31) *at position (x_j, t^n),*

$$\frac{w_j^{n+1} - w_j^n}{\Delta t} = w_t + \frac{\Delta t}{2} w_{tt} + O(\Delta t^2) \tag{32}$$

$$= w_t + \frac{\Delta t}{2} A^2 w_{xx} + O(\Delta t^2) \tag{33}$$

and

$$\frac{\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n - \widehat{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} = \frac{1}{2\Delta x} \widehat{A} \left(w_{j+1}^n - w_{j-1}^n \right) - \frac{\widehat{\alpha}}{2\Delta x} \left(w_{j-1}^n - 2w_j^n + w_{j+1}^n \right) \tag{34}$$

$$= \widehat{A}\, w_x - \frac{\widehat{\alpha} \Delta x}{2} w_{xx} + O(\Delta x^2). \tag{35}$$

Similarly,

$$\frac{\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^n - \widetilde{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} = \frac{1}{2\Delta x} \widetilde{A} \left(w_{j+1}^{n+1} - w_{j-1}^{n+1} \right) - \frac{\widetilde{\alpha}}{2\Delta x} \left(w_{j-1}^{n+1} - 2w_j^{n+1} + w_{j-1}^{n+1} \right) \tag{36}$$

$$= \widetilde{A} w_x(x_j, t^{n+1}) - \frac{\widetilde{\alpha} \Delta x}{2} w_{xx}(x_j, t^{n+1}) + O(\Delta x^2). \tag{37}$$

From (30),

$$w_x(x_j, t^{n+1}) = w_x(x_j, t^n) - \Delta t A w_{xx} + O(\Delta t^2), \tag{38}$$

while $w_{xx}(x_j, t^{n+1}) = w_{xx}(x_j, t^n) + O(\Delta t)$. Therefore,

$$\frac{\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^n - \widetilde{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} = \widetilde{A} w_x - \Delta t \widetilde{A} A w_{xx} - \frac{\widetilde{\alpha} \Delta x}{2} w_{xx} + O(\Delta x^2). \tag{39}$$

Now we plug (32), (34) *and* (39) *into* (23) *to obtain, always at position (x_j, t^n),*

$$\mathcal{I}_j^n(W^n, W^{n+1}) \tag{40}$$

$$= w_t + \frac{\Delta t}{2} A^2 w_{xx} + \widehat{A} w_x - \frac{\widehat{\alpha}}{2} \Delta x w_{xx} \tag{41}$$

$$+ \widetilde{A} w_x - \Delta t \widetilde{A} A w_{xx} - \frac{\widetilde{\alpha}}{2} \Delta x w_{xx} + O(\Delta x^2) \tag{42}$$

$$= w_t + (\widehat{A} + \widetilde{A}) w_x \tag{43}$$

$$+ \frac{\Delta t}{2} \left(A^2 - 2\widetilde{A} A - \widehat{\alpha} \frac{\Delta x}{\Delta t} \mathrm{Id} - \widetilde{\alpha} \frac{\Delta x}{\Delta t} \mathrm{Id} \right) w_{xx} + O(\Delta x^2). \tag{44}$$

This is $O(\Delta x^2)$ if and only if w fulfills (29) with

$$B = \frac{\Delta t}{2}\left(-A^2 + 2\widetilde{A}A + \widehat{\alpha}\frac{\Delta x}{\Delta t}\,\mathrm{Id} + \widetilde{\alpha}\frac{\Delta x}{\Delta t}\,\mathrm{Id}\right) \tag{45}$$

$$= \frac{\Delta t}{2}\left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t}\,\mathrm{Id} - (\widehat{A}-\widetilde{A})A\right). \tag{46}$$

Note that we have repeatedly used the assumption $O(\Delta x) = O(\Delta t)$. This proves the claim.

In the sequel, we show how (27) can be used to determine whether (and under what CFL condition) the IMEX scheme (23) is stable. We begin by deriving an exact solution to (29) using a Fourier ansatz. Note that A is a $d \times d$ matrix.

Lemma 2 *Let w_0 be given by*

$$w_0(x) = \sum_{k\in\mathbb{Z}} \begin{pmatrix} a_k^1 \\ \vdots \\ a_k^d \end{pmatrix} e^{i2\pi kx}. \tag{47}$$

Furthermore, let w be a solution to

$$w_t + Aw_x = Bw_{xx} \quad \forall (x,t) \in \Omega \times (0,T) \tag{48}$$

$$w(x,0) = w_0(x) \quad \forall x \in \Omega. \tag{49}$$

Then, w admits a representation

$$w(x,t) = \sum_{k\in\mathbb{Z}} \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} e^{i2\pi kx} \tag{50}$$

with $a_k^1, \dots, a_k^d$ fulfilling the system of d differential equations

$$\begin{pmatrix} a_k^1(t)' \\ \vdots \\ a_k^d(t)' \end{pmatrix} = \mathcal{A}_k \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} \tag{51}$$

for

$$\mathcal{A}_k := (-i2\pi kA - 4\pi^2 k^2 B) \tag{52}$$

and initial conditions

$$\begin{pmatrix} a_k^1(0) \\ \vdots \\ a_k^d(0) \end{pmatrix} = \begin{pmatrix} a_k^1 \\ \vdots \\ a_k^d \end{pmatrix}. \tag{53}$$

Proof The proof exploits direct computations and starts with assuming *that the representation* (50) *is correct. Thus, plugging* (50) *into* (48)*, one obtains*

$$\sum_{k\in\mathbb{Z}} \left(\begin{pmatrix} a_k^1(t)' \\ \vdots \\ a_k^d(t)' \end{pmatrix} + i2\pi k A \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} + 4\pi^2 k^2 B \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} \right) e^{i2\pi kx} = 0. \quad (54)$$

Exploiting the linear independence of $e^{i2\pi kx}$ *for different* k*, one obtains* (51)*.*

Remark 1 1. Every periodic piecewise smooth function w_0 can be written as in (47).
2. For future reference, we call $\mathcal{A}_k$ the *frequency matrices* of the modified equation (27).

The following corollary is a direct consequence from the theory of ordinary differential equations, and Plancherel's theorem.

Corollary 1 *We consider the setting as in Lemma 2. Then,*

$$\|w(\cdot,t)\|_{L^2(\Omega)} \leq \|w_0(\cdot)\|_{L^2(\Omega)} \quad (55)$$

holds if

$$\text{Real}(\mu_i) < 0 \quad (56)$$

for all eigenvalues μ_i *of* $\mathcal{A}_k$ *with* $k \in \mathbb{Z}\setminus\{0\}$*.*

Remark 2 One might argue that Corollary 1 is not needed in the sense that for every matrix B with B positive definite, there holds (55). Indeed, for B positive definite and, say A symmetric, one has, using an energy argument on equation (48) (integration is over Ω, and as the problem is periodic, the boundary terms vanish)

$$\frac{d}{dt}\frac{\|w\|^2_{L^2(\Omega)}}{2} = (w_t, w) = -(Aw_x, w) + (w, Bw_{xx}) \quad (57)$$

$$= (Aw, w_x) - (w_x, Bw_x) = ((wAw)_x, \frac{1}{2}) - (w_x, Bw_x) < 0 \quad (58)$$

and ergo (55). However, this is only a sufficient, not a necessary condition. Consider, e.g., the pair of matrices $A = \text{Id}$ and $B = \begin{pmatrix} 5 & 1 \\ -2 & 0 \end{pmatrix}$. Obviously, B is not positive definite (note that $x^T Bx < 0$ for, e.g., $x := (1, 10)^T$), however, the eigenvalues of $\mathcal{A}_k$ have negative real part, and consequently, the complete system (48) is stable. (A tedious computation reveals that the eigenvalues of $\mathcal{A}_k$ are $2\pi k\left((\pm\sqrt{17} - 5)\pi k - i\right)$.)

The real part of the eigenvalues of $\mathcal{A}_k$ is *not* affected by the terms coming from A if matrices A and B can be simultaneously diagonalized. This was the motivation for considering the characteristic splitting introduced in (12).

5 Stability of characteristic flux splittings

Now, we combine Theorem 1 and Corollary 1 to obtain a necessary criterion under what circumstances the IMEX scheme (23) is stable.

We consider the characteristic splitting (12) in the light of Corollary (1). For a generic splitting with *commuting* matrices $\widehat{A}$ and $\widetilde{A}$, the frequency matrix $\mathcal{A}_k$ can be written as

$$\mathcal{A}_k = -i2\pi kA - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t}\,\mathrm{Id} - (\widehat{A}-\widetilde{A})(\widehat{A}+\widetilde{A}) \right) \tag{59}$$

$$= -i2\pi kA - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t}\,\mathrm{Id} - \widehat{A}^2 + \widetilde{A}^2 \right). \tag{60}$$

Note that $\mathcal{A}_0 = 0$, since constant solutions of the modified equations are independent of time. Therefore we need to analyze only the case $k \neq 0$. As we rely on a characteristic splitting, all the matrices occurring in (60) can be written as $Q\Sigma Q^{-1}$ for some diagonal matrix Σ. Therefore, it is easy to see that the real part μ_i of the eigenvalues of $\mathcal{A}_k$ is given by

$$\mathrm{Real}(\mu_i) = 2\pi^2 k^2 \Delta t \left(-\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t} + \widehat{\lambda}_i^2 - \widetilde{\lambda}_i^2 \right) \tag{61}$$

where $\widehat{\lambda}_i$ and $\widetilde{\lambda}_i$ are eigenvalues to $\widehat{A}$ and $\widetilde{A}$, respectively. Claiming that $\mathrm{Real}(\mu_i)$ is negative leads to

$$\frac{(\widehat{\alpha}+\widetilde{\alpha})\,\Delta x}{\Delta t} > \widehat{\lambda}_i^2 - \widetilde{\lambda}_i^2. \tag{62}$$

This is a good result in the following sense: $\widehat{\lambda}_i^2$ can be bounded independently of ε. Therefore, for

$$\frac{\Delta t}{\Delta x} < \frac{\widehat{\alpha}+\widetilde{\alpha}}{\widehat{\lambda}_i^2}, \quad \forall i, \tag{63}$$

(62) and thus Corollary (1) holds. (63) is a restriction that can be made independent of ε. We summarize this in the following lemma.

Lemma 3 *The characteristic splitting as introduced in* (12) *can always be made stable with a time step size independent of* ε*.*

In the sequel, we consider a prototype system in more detail to obtain quantitative information.

5.1 Characteristic Splitting of prototype equation

In this section, we consider the prototype matrix from Section 2.1, as it allows for easy computations. The non-dimensionalized advective CFL number corresponding to A is denoted by

$$\widehat{\nu} := \frac{a\Delta t}{\Delta x}. \tag{64}$$

Given $k \neq 0$, we consider the frequency matrix $\mathcal{A}_k$ for the characteristic splitting introduced in (12). One potential advantage of the characteristic splitting is that one can compute the eigenvalues explicitly, as all the matrices commute:

Lemma 4 *The real part of the eigenvalues μ_i of $\mathcal{A}_k$ are given by*

$$\mathrm{Real}(\mu_1) = -2\pi^2 k^2 \Delta x(\widehat{\alpha} + \widetilde{\alpha}) + 2\pi^2 k^2 \Delta t a^2 \tag{65}$$

$$\mathrm{Real}(\mu_{2,3}) = \frac{-4\pi^2 k^2 \Delta t}{\varepsilon^2} + \frac{8\Delta t k^2 \pi^2}{\varepsilon} \tag{66}$$

$$+ 2\pi^2 k^2 a^2 \Delta t - 2\pi^2 k^2 \Delta x(\widehat{\alpha} + \widetilde{\alpha}) \pm 4\sqrt{2} a \pi^2 k^2 \Delta t. \tag{67}$$

Proof With the notation introduced in (10) *and* (12)*, see also* (60)*, there holds*

$$\mathcal{A}_k = -i2\pi k Q \Lambda Q^{-1} - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha} + \widetilde{\alpha})\,\Delta x}{\Delta t} \,\mathrm{Id} - (Q(\widehat{\Lambda} - \widetilde{\Lambda})Q^{-1}) Q \Lambda Q^{-1} \right) \tag{68}$$

$$= Q \left(-i2\pi k \Lambda - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha} + \widetilde{\alpha})\,\Delta x}{\Delta t} \,\mathrm{Id} - (\widehat{\Lambda} - \widetilde{\Lambda})\Lambda \right) \right) Q^{-1} \tag{69}$$

$$= Q \,\mathrm{diag} \begin{pmatrix} -i2\pi k(a - \frac{\sqrt{2}}{\varepsilon}) - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\Delta x}{\Delta t} + \frac{2}{\varepsilon^2} - \frac{4}{\varepsilon} + 2\sqrt{2}a - a^2 \right) \\ -i2\pi k a - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\Delta x}{\Delta t} - a^2 \right) \\ -i2\pi k(a + \frac{\sqrt{2}}{\varepsilon}) - 2\pi^2 k^2 \Delta t \left(\frac{(\widehat{\alpha}+\widetilde{\alpha})\Delta x}{\Delta t} + \frac{2}{\varepsilon^2} - \frac{4}{\varepsilon} - 2\sqrt{2}a - a^2 \right) \end{pmatrix} Q^{-1} \tag{70}$$

Thus, one can conclude that the eigenvalues of $\mathcal{A}_k$ are those of the diagonal matrix. Sorting the eigenvalues conveniently, starting with the one in the middle, one can conclude that their Real parts are given by formulae (65) *and* (66)*.*

The problem under consideration has two asymptotics, namely the one associated to $\varepsilon \to 0$, and the other one associated to $\Delta t \to 0$ (which automatically includes $\Delta x \to 0$). We immediately obtain the following condition for the negativity of the first eigenvalue:

Lemma 5 $\mathrm{Real}(\mu_1) < 0$ *under the* CFL *condition*

$$\widehat{\nu} < \nu_1 := \frac{\widehat{\alpha} + \widetilde{\alpha}}{a}. \tag{71}$$

Remark 3 Condition (71) is a condition for the nonstiff CFL number $\widehat{\nu}$ from (64). It depends on both $\widehat{\alpha}$ and $\widetilde{\alpha}$. The coefficient $\widehat{\alpha}$ encodes the upwind viscosity of the explicit numerical flux (24), and is usually chosen as $a+\sqrt{2}$, the largest eigenvalue of the nonstiff matrix. There is more freedom to choose the viscosity coefficient of the implicit numerical flux (25), and limiting choices are either $\widetilde{\alpha} = \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}$ (the largest eigenvalue of the nonstiff matrix) or $\widetilde{\alpha} = 0$. In both cases, $\widehat{\alpha}+\widetilde{\alpha} \geq a+\sqrt{2}$, which gives the sufficient stability condition

$$\widehat{\nu} < \frac{a+\sqrt{2}}{a}. \tag{72}$$

This is independent of ε.

Now we discuss $\mathrm{Real}(\mu_{2,3})$. Obviously, for Δt fixed and $\varepsilon \to 0$, $\mathrm{Real}(\mu_{2,3}) < 0$, which directly yields the following Theorem:

Theorem 2 *Let Δt be fixed. Then, there exists an $\varepsilon_0 > 0$, such that for all $\varepsilon < \varepsilon_0$ and all $k \neq 0$, $\mathrm{Real}(\mu_{2,3}) < 0$.*

However, this is not the full asymptotics. We therefore change the point of view: Given a *fixed* ε, for which Δt is $\mathrm{Real}(\mu_{2,3}) < 0$? The following lemma provides the crucial estimate:

Lemma 6 *We define*

$$\varphi(a) := \frac{\sqrt{2}}{a+2\sqrt{2}}. \tag{73}$$

Now, consider two cases as follows:

1. *Let $\varepsilon \leq \varphi(a)$. Then, $\mathrm{Real}(\mu_{2,3}) < 0$ holds unconditionally.*
2. *Let $\varphi(a) < \varepsilon$. Then, $\mathrm{Real}(\mu_{2,3}) < 0$ holds for*

$$\widehat{\nu} < \frac{(\widehat{\alpha}+\widetilde{\alpha})a}{\frac{-2+4\varepsilon}{\varepsilon^2} + a^2 + 2\sqrt{2}a} \tag{74}$$

which is automatically fulfilled for

$$\widehat{\nu} \leq \frac{(\widehat{\alpha}+\widetilde{\alpha})a}{(a+\sqrt{2})^2}. \tag{75}$$

Proof Obviously, one only has to show that

$$0 > \frac{-4\pi^2k^2\Delta t}{\varepsilon^2} + \frac{8\Delta t k^2\pi^2}{\varepsilon} + 2\pi^2k^2a^2\Delta t - 2\pi^2k^2\Delta x(\widehat{\alpha}+\widetilde{\alpha}) + 4\sqrt{2}a\pi^2k^2\Delta t. \tag{76}$$

We substitute $\Delta x = \frac{a\Delta t}{\widehat{\nu}}$ and obtain

$$0 > \frac{-4}{\varepsilon^2} + \frac{8}{\varepsilon} + 2a^2 - \frac{2(\widehat{\alpha}+\widetilde{\alpha})a}{\widehat{\nu}} + 4\sqrt{2}a \tag{77}$$

$$\Leftrightarrow \frac{(\widehat{\alpha}+\widetilde{\alpha})a}{\widehat{\nu}} > \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a. \tag{78}$$

(78) *is trivially fulfilled, if the right-hand side is not positive, i.e., if*

$$0 \geq \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a \tag{79}$$

$$\Leftrightarrow 0 \geq -2 + 4\varepsilon + \varepsilon^2 \left(a^2 + 2\sqrt{2}a\right) \tag{80}$$

$$\Leftrightarrow 0 \leq \varepsilon \leq \varphi(a). \tag{81}$$

This proves the first claim that for all $\varepsilon \leq \varphi(a)$*, both* $\mathrm{Real}(\mu_{2,3})$ *are negative.*

Now let $\varphi(a) < \varepsilon$*. In this case, the right-hand side of* (78) *is positive, and therefore one has a restriction on* $\widehat{\nu}$*. One can compute*

$$\frac{(\widehat{\alpha} + \widetilde{\alpha})a}{\widehat{\nu}} > \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a \tag{82}$$

$$\Leftrightarrow \frac{\widehat{\nu}}{(\widehat{\alpha} + \widetilde{\alpha})a} < \frac{1}{\frac{-2+4\varepsilon}{\varepsilon^2} + a^2 + 2\sqrt{2}a}. \tag{83}$$

As $\frac{-2+4\varepsilon}{\varepsilon^2} \leq 2$ *for* $0 \leq \varepsilon \leq 1$*, this is fulfilled if*

$$\frac{\widehat{\nu}}{(\widehat{\alpha} + \widetilde{\alpha})a} \leq \frac{1}{2 + a^2 + 2\sqrt{2}a} = \frac{1}{(a + \sqrt{2})^2} \tag{84}$$

This proves the lemma.

With Lemmas 5 and 6 we obtain the following

Theorem 3 *For* $k \neq 0$*,* $\mathrm{Real}(\mu_1) < 0$ *and* $\mathrm{Real}(\mu_{2,3}) < 0$ *if*

$$\widehat{\nu} < \nu_2 := \nu_1 \, \psi(\varepsilon) \tag{85}$$

with

$$\psi(\varepsilon) := \psi(\varepsilon, a) := \begin{cases} \min\left(1, \frac{a^2}{(\sqrt{2}+a)^2 - 2\left(\frac{1-\varepsilon}{\varepsilon}\right)^2}\right), & \varepsilon > \varphi(a) \\ 1, & \varepsilon \leq \varphi(a). \end{cases} \tag{86}$$

From the previous considerations, we can conclude that our scheme is stable under (at most!) the nonstiff CFL condition.

Corollary 2 *We choose* $\widehat{\alpha} = a + \sqrt{2}$*. Then, the scheme is stable for all* $0 \leq \varepsilon \leq 1$ *if*

$$\frac{(a + \sqrt{2})\Delta t}{\Delta x} < 1. \tag{87}$$

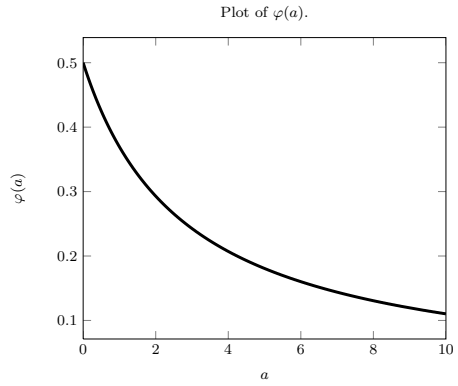

Fig. 1 Plot of function φ from (73).

Proof Consider expression (85) *and plug in the definition of both* ψ *from* (86) *and* ν_1 *from* (71)*. We consider case* $\varepsilon > \varphi(a)$ *first, and obtain*

$$\left(\frac{\widehat{\alpha}+\widetilde{\alpha}}{a}\right)\min\left(1,\frac{a^2}{(\sqrt{2}+a)^2-2\left(\frac{1-\varepsilon}{\varepsilon}\right)^2}\right) \geq \left(\frac{\widehat{\alpha}+\widetilde{\alpha}}{a}\right)\left(\frac{a^2}{(\sqrt{2}+a)^2}\right) \tag{88}$$

$$\geq \frac{(a+\sqrt{2})a}{(\sqrt{2}+a)^2} \geq \frac{a}{a+\sqrt{2}}. \tag{89}$$

Ergo, $\widehat{\nu} < \frac{a}{a+\sqrt{2}}$ *is sufficient for* (85)*, which implies* (87)*. Similarly, for* $\varepsilon \leq \varphi(a)$*, one can easily show that* $\widehat{\nu} < \nu_1$ *is fulfilled given that* (87) *holds.*

Remark 4 – The function $\varphi(a)$ has been plotted in Fig. 1. Note that the bound $\varepsilon \leq \varphi(a)$ implies that there is an $\varepsilon_0 > 0$, such that $\varepsilon < \varepsilon_0$. This is particularly interesting if one is only interested in the case $\varepsilon \to 0$, because ultimately, one only has to respect CFL condition (71) which is a condition on the convective CFL number only.

– In Fig. 2, we plotted the numerically determined maximum allowable $\widehat{\nu}$ values such that the real part of the eigenvalues of $\mathcal{A}_k$ are negative. For the particular computations, we choose $a = 2, \Delta x = 10^{-2}$, $\Delta t = \frac{\Delta x}{a\widehat{\nu}}$, $\widehat{\alpha} = 2+\sqrt{2}$, $\widetilde{\alpha} = 0$ or $\widetilde{\alpha} = \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}$ and determine the maximum $\widehat{\nu}$, such that $\mathrm{Real}(\mu_i) < 0$ for all $i = 1, 2, 3$ and for all $|n| \leq 25$. (Note however from our description of the eigenvalues in (65)-(66), the results are independent of Δt, Δx and n.)

6 On the instability of splittings for the Euler equations

The approach pursued in this paper for the simple splitting as introduced in (12) can also be applied to other, more involved splittings. However, computing the eigenvalues of $\mathcal{A}_k$ analytically may be hard or even impossible, because

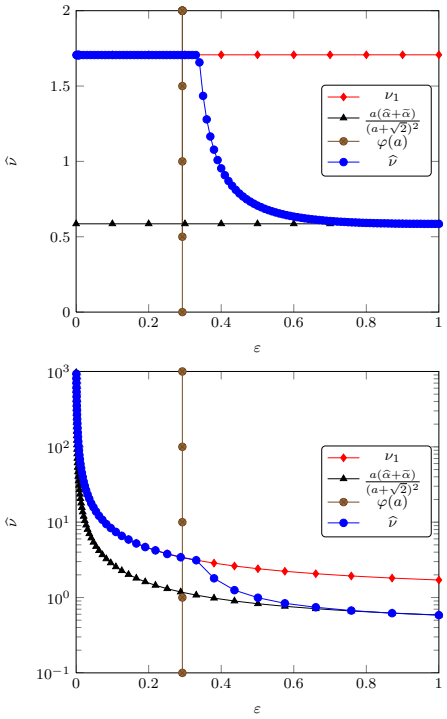

Fig. 2 Determined CFL number versus a-priori estimates. Left: $\widetilde{\alpha} = 0$, Right: $\widetilde{\alpha} = \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}$.

for general splittings, there is no simultaneous diagonalization of the matrices involved and ergo, one can not compute the eigenvalues easily. What is, however, possible, is to numerically evaluate, for different ratios of Δx and Δt, the eigenvalues of $\mathcal{A}_k$ and decide whether their real parts are negative. Of course this can only be performed for a finite number of Fourier modes k, however, it gives an idea of what is to be expected.

We consider equation (3) with the usual equation of state for pressure p,

$$p = (\gamma - 1)(E - \frac{1}{2}\rho u^2), \tag{90}$$

linearized around a state $v_0 := (\rho_0, \rho_0 u_0, E_0)$. For the linearization matrix $f'(v_0)$, we consider two different splittings:

1. The first splitting is the characteristic splitting introduced in (12) using diagonal matrices $\widehat{\Lambda}$ and $\widetilde{\Lambda}$ as given in (19)-(20).
2. The second one is a splitting taken from literature [14]. This splitting is a modification of Klein's original splitting [12], and it is given by a splitting

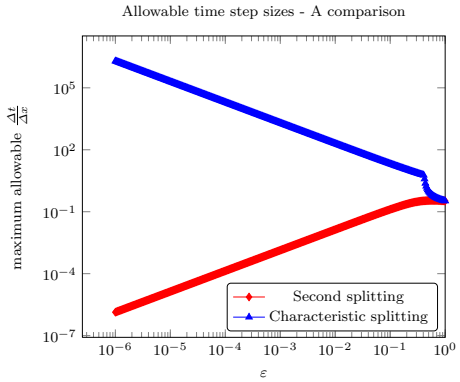

Fig. 3 Comparison of classical versus characteristic splitting

of the flux function $f(v)$ into the sum of

$$\widehat{f}(v) := \left(\rho u, \rho u^2 + p, u(E + \Pi)\right)^T, \tag{91}$$

$$\widetilde{f}(v) := \left(0, \frac{1-\varepsilon^2}{\varepsilon^2} p, u(p - \Pi)\right). \tag{92}$$

Π is an auxiliary pressure function, and it is defined by

$$\Pi(x,t) := \varepsilon^2 p(x,t) + (1-\varepsilon^2) \inf_x p(x,t). \tag{93}$$

The splittings are linearized, and one obtains an admissible splitting of $f'(v_0)$.

We perform the following numerical experiment: For the values $\rho_0 = 1$, $p_0 = 1$, and $u_0 = \frac{\sqrt{\gamma}}{2}$, we compute, for $\Delta x = 10^{-2}$, the largest Δt such that the eigenvalues of the matrix $\mathcal{A}_k$ have negative real part. The numerical viscosities are chosen as the maximum eigenvalues of the corresponding matrices. Results can be seen in Fig. 3. It can be clearly seen that, in order to be stable, the splitting taken from literature [14] needs a time step that decreases with ε. This is clearly not what one focuses on in asymptotic preserving schemes; it has however been already numerically experienced by the authors, which is why they include a stabilization term to render the scheme stable.

7 Conclusions and Outlook

We developed a technique to investigate the stability and the largest allowable time steps for low-order IMEX schemes based on a general class of splittings for linear hyperbolic conservation laws. Our analysis confirms the common believe that the nonstiff CFL number plays a crucial role in the stability condition. Indeed, it needs to be smaller than a constant multiple of the numerical

viscosities, and in the examples we considered, it is independent of the small parameter ε.

Furthermore, we introduced a new way of obtaining suitable splittings via characteristic decomposition of the flux Jacobian. We demonstrated that the splitting does influence the stability of the resulting method, and therefore, one should put effort into designing suitable flux splittings.

The extension of this analysis to *nonlinear* systems of conservation laws is not straightforward. However, considering the linearized equations, the analysis can be easily used as a guiding principle, similar as the von-Neumann analysis. Furthermore, the extension of both the stability analysis and the characteristic splittings to multiple dimensions is not straightforward, as multiple matrices do not necessarily commute. However, as the numerical flux only acts on the flux function in normal direction, we believe that the construction of a characteristic splitting is, nevertheless, possible.

The results presented in this paper do not include any statements about the quality of the approximation of the IMEX scheme, but we obtain a pure stability analysis. Future work will account for such issues in the context of compressible flows.

References

1. Arun, K., Noelle, S.: An asymptotic preserving scheme for low froude number shallow flows. IGPM Preprint 352 (2012)
2. Ascher, U.M., Ruuth, S.J., Spiteri, R.J.: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Applied Numerical Mathematics **25**, 151–167 (1997)
3. Boscarino, S.: Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems. SIAM Journal on Numerical Analysis **45**, 1600–1621 (2007)
4. Cordier, F., Degond, P., Kumbaro, A.: An asymptotic-preserving all-speed scheme for the euler and navier-stokes equations. Journal of Computational Physics **231**, 5685–5704 (2012)
5. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Mathematische Annalen **100**(1), 32–74 (1928). URL http://dx.doi.org/10.1007/BF01448839
6. Degond, P., Lozinski, A., Narski, J., Negulescu, C.: An asymptotic-preserving method for highly anisotropic elliptic equations based on a micro-macro decomposition. Journal of Computational Physics **231**, 2724–2740 (2012)
7. Degond, P., Tang, M.: All speed scheme for the low mach number limit of the isentropic euler equation. Communications in Computational Physics **10**, 1–31 (2011)
8. Haack, J., Jin, S., Liu, J.G.: An all-speed asymptotic-preserving method for the isentropic euler and navier-stokes equations. Communications in Computational Physics **12**, 955–980 (2012)
9. Jin, S.: Efficient asymptotic-preserving (ap) schemes for some multiscale kinetic equations. SIAM Journal on Scientific Computing **21**, 441–454 (1999)
10. Jin, S.: Asymptotic preserving (ap) schemes for multiscale kinetic and hyperbolic equations: A review. Riv. Mat. Univ. Parma **3**, 177–216 (2012)
11. Jin, S., Pareschi, L., Toscani, G.: Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. SIAM Journal on Numerical Analysis **35**, 2405–2439 (1998)
12. Klein, R.: Semi-Implicit Extension of a Godunov-Type Scheme Based on Low Mach Number Asymptotics I: One-Dimensional Flow. Journal of Computational Physics **121**, 213–237 (1995)

13. Lax, P.: On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients. Communications on Pure and Applied Mathematics **14**, 497–520 (1961)
14. Noelle, S., Bispen, G., Arun, K., Lukacova-Medvidova, M., Munz, C.D.: An asymptotic preserving all mach number scheme for the euler equations of gas dynamics. IGPM Preprint 348 (2012)
15. Russo, G., Boscarino, S.: IMEX Runge-Kutta schemes for hyperbolic systems with diffusive relaxation. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012) (2012)
16. Schütz, J.: An asymptotic preserving method for linear systems of balance laws based on Galerkin's method. Journal of Scientific Computing (2013). DOI 10.1007/s10915-013-9801-1
17. Strang, G.: Accurate partial difference methods. Numerische Mathematik **6**, 37–46 (1964)
18. Warming, R., Hyett, B.J.: The modified equation approach to the stability and accuracy of finite-difference methods. Journal of Computational Physics **14**, 159–179 (1974)