

# Flux Splitting for stiff equations: A notion on stability

Jochen Schütz · Sebastian Noelle

Received: date / Accepted: date

**Abstract** For low Mach number flows, there is a strong recent interest in the development and analysis of *IMEX* (implicit/explicit) schemes, which rely on a splitting of the convective flux into stiff and nonstiff parts. A key ingredient of the analysis is the so-called *Asymptotic Preserving* (AP) property, which guarantees uniform consistency and stability as the Mach number goes to zero. While many authors have focussed on asymptotic consistency, we study asymptotic stability in this paper: does an IMEX scheme allow for a CFL number which is independent of the Mach number? We derive a stability criterion for a general linear hyperbolic system. In the decisive eigenvalue analysis, the advective term, the upwind diffusion and a quadratic term stemming from the truncation in time all interact in a subtle way. As an application, we show that a new class of splittings based on characteristic decomposition, for which the commutator vanishes, avoids the deterioration of the time step which has sometimes been observed in the literature.

**Keywords:** IMEX Finite Volume, Asymptotic Preserving, Flux Splitting, Modified Equation, Stability Analysis

**AMS subject classification:** 35L65, 76M45, 65M08

## 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 one is interested in computing waves with vastly different speeds. The goal is to resolve slow waves accurately and

---

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

efficiently with a large time step, while approximating the fast waves in a stable way, using the same time step.

There is vast literature on the computation of low-speed viscous and inviscid fluid flows. Arguably the first contribution within this field is Chorin's algorithm [5], who proposes to solve the incompressible Navier-Stokes equations using a projection method. Similar methods have also been used in, e.g., [6]. A different approach to reduce the stiffness occurring at low Mach numbers is to introduce preconditioning, i.e., to multiply the temporal derivative with a suitable matrix, see the pioneering work by Turkel [32], and, built on this result, the works by Guillard et al. [14, 15, 26]. In [15], the author identifies the main problem in approximating low-speed flows: Roughly speaking, the variation of pressure is of second order in the Mach number  $Ma$ . However, 'traditional' Godunov schemes tend to produce pressure variations that are of first order in the Mach number, thus for  $Ma \rightarrow 0$ , there can be no uniform convergence. For an extensive additional analysis, in particular with respect to suitable initial conditions, we refer to Dellacherie [12]. We do not intend to give a fully exhaustive overview on this topic. For an overview on the treatment of low-speed flows, we refer to [4] and the references therein; a more recent survey was given in [23].

A class of algorithms that has found particular attention are the *Asymptotic Preserving* schemes introduced by Jin [18], built on work with Pareschi and Toscani [20]. For an excellent review article, consult [19]; we refer to [1, 7, 10, 11, 16, 27, 30] for various applications of this method in different contexts.

Many algorithms, especially those used within the *Asymptotic Preserving* schemes rely on identifying *stiff* and *nonstiff* parts of the underlying equation. This point is generally considered crucial, and the hope is that 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 [22].

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]. We refer to [3, 29] for an interesting discussion on the quality of these schemes in the asymptotic limit.

As far as the authors can see, a fully nonlinear asymptotic stability analysis for the non-isentropic Euler equations is still out of reach. In this work, we attempt to reveal an important structural stability property of flux splittings via the considerably simpler *modified equation* analysis [33] for a prototype,  $3 \times 3$  linear system of conservation laws.

More specifically we derive the modified parabolic system of equations of second order and investigate under what conditions its solutions are bounded in the  $L^2$ -norm. For simple problems, one can investigate these conditions analytically. This approach is closely related to the classical *von-Neumann* stability analysis (see, e.g., [25, 28]). Strang [31] 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 at first sight.

Throughout the paper, we consider the linear hyperbolic system of conservation laws

$$u_t + Au_x = 0 \quad \forall (x, t) \in \Omega \times (0, T) \quad (1a)$$

$$u(x, 0) = u_0(x) \quad \forall x \in \Omega \quad (1b)$$

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 set  $\Omega := [0, 1]$  and consider smooth periodic solutions  $u$  with initial data  $u_0$ . Furthermore, we assume that the matrix  $A$  is a function of a parameter  $\varepsilon \in (0, 1]$  such that with  $\varepsilon \rightarrow 0$ , some eigenvalues of  $A$  diverge towards infinity.

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

$$v_t + f(v)_x = 0, \quad (2)$$

e.g., the (non-dimensionalized) Euler equations at low Mach number  $\varepsilon$ , with characteristic quantities density, momentum and total energy,  $v = (\rho, \rho v, E)^T$ , and

$$f(\rho, \rho v, E) := \left( \rho v, \rho v^2 + \frac{p}{\varepsilon^2}, v(E + p) \right)^T. \quad (3)$$

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

$$\lambda = v, v \pm \frac{c}{\varepsilon},$$

where  $c := \sqrt{\frac{\gamma p}{\rho}}$  denotes the speed of sound.  $\gamma$  is the ratio of specific heats, frequently taken to be  $\gamma = 1.4$  for air. Note that two eigenvalues tend to infinity as  $\varepsilon \rightarrow 0$ .

To highlight the difficulties posed by eigenvalues of multiple scales we briefly discuss a standard, explicit finite volume scheme for (2)

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

with consistent numerical flux  $\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n$ . From the stability conditions by Courant, Friedrichs and Lewy [8], it is known that explicit schemes are only stable under a CFL condition, which is typically given by

$$\nu_{\max} := \lambda_{\max} \frac{\Delta t}{\Delta x} = \frac{(v + \frac{c}{\varepsilon}) \Delta t}{\Delta x} < 1. \quad (4)$$

In the limit as  $\varepsilon \rightarrow 0$ , one mainly wants to resolve the advective wave traveling with speed  $v$ . Given the restrictive CFL condition (4), this would imply that one needs  $\mathcal{O}(\varepsilon^{-1})$  steps to advect a signal across a single grid cell. For small

$\varepsilon$ , this is prohibitively inefficient, and for many schemes also prohibitively dissipative. However, using

$$\hat{\nu} := v \frac{\Delta t}{\Delta x} < 1 \quad (5)$$

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 the flux  $f$  into components with 'slow' and 'fast' waves. More precisely, in the context of (1), one seeks matrices  $\hat{A}$  and  $\tilde{A}$ , such that

$$A = \hat{A} + \tilde{A}, \quad (6)$$

with the following conditions posed on  $\hat{A}$  and  $\tilde{A}$ :

**Definition 1** The splitting (6) is called *admissible*, if

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

In this work, we give a recipe for identifying stable classes of flux splittings for the linear hyperbolic system (1). We use the well-known modified equation analysis as a tool for (heuristically) investigating (linear)  $L^2$ -stability.

The paper is outlined as follows: In Section 2, we introduce the lowest-order IMEX scheme, while in Section 3, we investigate this scheme using the modified equation approach. In Section 4, we introduce so-called characteristic splittings, which are a basic ingredient for a uniformly stable scheme. Based on those sections, we show our main result in Section 5, Theorem 2: Characteristic splittings are stable in the sense as explained in Section 3 with a time step size *independently* of  $\varepsilon$ . Section 6 shows an example of a scheme that is only stable with a time step size decreasing with  $\varepsilon$ , i.e., the allowable  $\Delta t$  such that the scheme is stable behaves as  $\Delta t \propto \varepsilon$ , which, for small  $\varepsilon$  is obviously not the desired effect. In Section 7, we apply our analysis to the linearized Euler equations and show some numerical results that substantiate the theory developed in this paper. Finally, Section 8 offers conclusions and possible future work.

## 2 IMEX Discretization

Based on a splitting as given in (6), we introduce a straightforward first-order IMEX discretization for (1) based on nonstiff and stiff numerical fluxes  $\hat{\mathcal{H}}$  and  $\tilde{\mathcal{H}}$ . We assume that the temporal domain is subdivided as

$$0 = t^0 < t^1 < \dots < t^N = T$$

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

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

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 a numerical approximation to  $u(x_j, t^n)$  by  $u_j^n$ . Furthermore, the vector  $(u_0^n, \dots, u_J^n)$  is denoted by  $\mathbf{U}^n$ . Now we can introduce a (classical) first-order IMEX scheme:

**Definition 2** A sequence  $\mathbf{U} = (\mathbf{U}^0, \dots, \mathbf{U}^N)$  is a solution to an IMEX discretization, given that

$$\mathcal{I}_j^n(\mathbf{U}) = 0, \forall j \in \{0, \dots, J\}, \forall n \in \{0, \dots, N-1\},$$

where

$$\mathcal{I}_j^n(\mathbf{U}) := \frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n - \widehat{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} + \frac{\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} - \widetilde{\mathcal{H}}_{j-\frac{1}{2}}^{n+1}}{\Delta x}. \quad (7)$$

Here, nonstiff and stiff numerical fluxes are defined by

$$\widehat{\mathcal{H}}_{j+\frac{1}{2}}^n := \frac{1}{2} \widehat{A} (u_{j+1}^n + u_j^n) - \frac{\widehat{\alpha}}{2} (u_{j+1}^n - u_j^n) \quad (8)$$

$$\widetilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} := \frac{1}{2} \widetilde{A} (u_{j+1}^{n+1} + u_j^{n+1}) - \frac{\widetilde{\alpha}}{2} (u_{j+1}^{n+1} - u_j^{n+1}), \quad (9)$$

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

*Remark 1*  $\widehat{\mathcal{H}}$  and  $\widetilde{\mathcal{H}}$  are given in the so-called viscosity form of a numerical flux, see, e.g., [13]. More generally, one can also consider matrices for  $\widehat{\alpha}$  and  $\widetilde{\alpha}$  instead of scalars. This plays a role in preconditioned schemes; here, it is omitted for the sake of simplicity.

For fixed  $\varepsilon$ , consistency analysis of the scheme is well-known, see, e.g., [9, 13, 24]. However, we have to consider both  $\varepsilon$  and  $\Delta t$  as small parameters. The crucial point is that we restrict our analysis to cases where the magnitude of  $u$  and its first and second derivatives are independent of  $\varepsilon$ . (Especially, no derivative behaves as  $O(\varepsilon^{-1})$  or worse.) This assumption is reasonable, as only those solutions allow for an asymptotic limit as  $\varepsilon \rightarrow 0$ . Similar assumptions have been made in [21].

**Lemma 1** Let  $\underline{\mathbf{U}}^n$  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  $\varepsilon$ ; and let  $\underline{\mathbf{U}} := (\underline{\mathbf{U}}^0, \dots, \underline{\mathbf{U}}^N)$ . Furthermore, let  $O(\Delta x) = O(\Delta t)$ . Then, the local truncation error is of order  $\Delta x$ , i.e., there holds for all  $j = \{0, \dots, J\}$  and for all  $n \in \{0, \dots, N-1\}$ ,

$$\mathcal{I}_j^n(\underline{\mathbf{U}}) = O(\Delta x).$$

*Proof* Apply a Taylor expansion to (7) and note that all the derivatives of  $u$  can be bounded independently of  $\varepsilon$ .  $\square$

*Remark 2* Note that the condition that the derivatives of  $u$  can be bounded *independently* of  $\varepsilon$  is indeed a condition on the initial datum  $u_0$ . Not every initial data gives rise to such a solution. More precisely, this means that as  $\varepsilon \rightarrow 0$ , there is indeed a solution to (1) that does not blow up. In the context of the low-Froude or low-Mach number limit, such a choice of initial conditions is often called *well-prepared initial data*. For a discussion of the influence of initial conditions on the limit solution, we refer to the work by Klainerman and Majda [21]. However, also for examples from ordinary differential equations, there are analogues, see, e.g., [3, 17].

### 3 Modified Equation Analysis

In this section, we derive the modified equation [33] corresponding to (7). 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 (7).

**Theorem 1** *Let  $w$  be a smooth solution of*

$$w_t + Aw_x = \frac{\Delta t}{2} \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - (\hat{A} - \tilde{A})A \right) w_{xx}. \quad (10)$$

*Furthermore, we consider vectors  $\underline{\mathbf{W}}^n := (w(x_0, t^n), \dots, w(x_J, t^n))$  and  $\underline{\mathbf{W}} := (\underline{\mathbf{W}}^0, \dots, \underline{\mathbf{W}}^N)$ . Then, for fixed  $\varepsilon$  and  $O(\Delta x) = O(\Delta t)$ , the IMEX scheme (7) is a second order accurate discretization of (10), i.e.*

$$\mathcal{I}_j^n(\underline{\mathbf{W}}) = O(\Delta x^2).$$

*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

$$w_t + Aw_x = Bw_{xx} \quad (11)$$

for a (yet unknown) viscosity matrix  $B$  that is in class  $O(\Delta x)$ . Using (11), one can conclude that

$$w_t = -Aw_x + Bw_{xx} \quad (12)$$

$$w_{tt} = -A(w_t)_x + B(w_t)_{xx} \stackrel{(12)}{=} A^2 w_{xx} + O(\Delta x). \quad (13)$$

Note that this holds due to  $O(\Delta x) = O(\Delta t)$ , which we will from now on exploit frequently. To simplify the presentation, we slightly abuse our notation, and

write  $\underline{w}_j^n$  for  $w(x_j, t^n)$ . Using (13) at position  $(x_j, t^n)$ ,

$$\begin{aligned}\frac{\underline{w}_j^{n+1} - \underline{w}_j^n}{\Delta t} &= w_t + \frac{\Delta t}{2} w_{tt} + O(\Delta x^2) \\ &= w_t + \frac{\Delta t}{2} A^2 w_{xx} + O(\Delta x^2)\end{aligned}\quad (14)$$

and

$$\begin{aligned}\frac{\hat{\mathcal{H}}_{j+\frac{1}{2}}^n - \hat{\mathcal{H}}_{j-\frac{1}{2}}^n}{\Delta x} &= \frac{1}{2\Delta x} \hat{A} (\underline{w}_{j+1}^n - \underline{w}_{j-1}^n) - \frac{\hat{\alpha}}{2\Delta x} (\underline{w}_{j-1}^n - 2\underline{w}_j^n + \underline{w}_{j+1}^n) \\ &= \hat{A} w_x - \frac{\hat{\alpha}\Delta x}{2} w_{xx} + O(\Delta x^2).\end{aligned}\quad (15)$$

Similarly,

$$\begin{aligned}\frac{\tilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} - \tilde{\mathcal{H}}_{j-\frac{1}{2}}^{n+1}}{\Delta x} &= \frac{1}{2\Delta x} \tilde{A} (\underline{w}_{j+1}^{n+1} - \underline{w}_{j-1}^{n+1}) - \frac{\tilde{\alpha}}{2\Delta x} (\underline{w}_{j-1}^{n+1} - 2\underline{w}_j^{n+1} + \underline{w}_{j+1}^{n+1}) \\ &= \tilde{A} w_x(x_j, t^{n+1}) - \frac{\tilde{\alpha}\Delta x}{2} w_{xx}(x_j, t^{n+1}) + O(\Delta x^2).\end{aligned}$$

From (12),

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

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

$$\frac{\tilde{\mathcal{H}}_{j+\frac{1}{2}}^{n+1} - \tilde{\mathcal{H}}_{j-\frac{1}{2}}^{n+1}}{\Delta x} = \tilde{A} w_x - \Delta t \tilde{A} A w_{xx} - \frac{\tilde{\alpha}\Delta x}{2} w_{xx} + O(\Delta x^2). \quad (16)$$

Now we plug (14), (15) and (16) into (7) to obtain, always at position  $(x_j, t^n)$ ,

$$\begin{aligned}\mathcal{I}_j^n(\mathbf{W}) &= w_t + \frac{\Delta t}{2} A^2 w_{xx} + \hat{A} w_x - \frac{\hat{\alpha}}{2} \Delta x w_{xx} \\ &\quad + \tilde{A} w_x - \Delta t \tilde{A} A w_{xx} - \frac{\tilde{\alpha}}{2} \Delta x w_{xx} + O(\Delta x^2) \\ &= w_t + (\hat{A} + \tilde{A}) w_x \\ &\quad + \frac{\Delta t}{2} \left( A^2 - 2\tilde{A}A - \hat{\alpha} \frac{\Delta x}{\Delta t} \text{Id} - \tilde{\alpha} \frac{\Delta x}{\Delta t} \text{Id} \right) w_{xx} + O(\Delta x^2).\end{aligned}$$

This is  $O(\Delta x^2)$  if and only if  $w$  fulfills (11) with

$$\begin{aligned}B &= \frac{\Delta t}{2} \left( -A^2 + 2\tilde{A}A + \hat{\alpha} \frac{\Delta x}{\Delta t} \text{Id} + \tilde{\alpha} \frac{\Delta x}{\Delta t} \text{Id} \right) \\ &= \frac{\Delta t}{2} \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - (\hat{A} - \tilde{A})A \right).\end{aligned}\quad (17)$$

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

In the sequel, we show how (10) can be used to determine a necessary condition under what CFL condition the IMEX scheme (7) is stable. We begin by deriving an exact solution to (11) using a Fourier ansatz. Note that  $A$  is a  $d \times d$  matrix.

**Lemma 2** *Let  $w_0$  be given by*

$$w_0(x) = \sum_{k \in \mathbb{Z}} \begin{pmatrix} a_{0k}^1 \\ \vdots \\ a_{0k}^d \end{pmatrix} e^{i2\pi kx}. \quad (18)$$

*Furthermore, let  $w$  be a solution to*

$$\begin{aligned} w_t + Aw_x &= Bw_{xx} \quad \forall (x, t) \in \Omega \times (0, T) \\ w(x, 0) &= w_0(x) \quad \forall x \in \Omega. \end{aligned} \quad (19)$$

*Then,  $w$  admits a representation*

$$w(x, t) = \sum_{k \in \mathbb{Z}} \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} e^{i2\pi kx} \quad (20)$$

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

$$\begin{pmatrix} a_k^1(t)' \\ \vdots \\ a_k^d(t)' \end{pmatrix} = \mathcal{A}_k \begin{pmatrix} a_k^1(t) \\ \vdots \\ a_k^d(t) \end{pmatrix} \quad (21)$$

*for*

$$\mathcal{A}_k := (-i2\pi kA - 4\pi^2 k^2 B) \quad (22)$$

*and initial conditions*

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

*Proof* The proof exploits direct computations and starts with *assuming* that the representation (20) is correct. Thus, plugging (20) into (19), one obtains

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

Exploiting the linear independence of  $e^{i2\pi kx}$  for different  $k$ , one obtains (21).  $\square$



*Remark 3* – Every periodic smooth function  $w_0$  can be written as in (18).  
 – For future reference, we call  $\mathcal{A}_k$  the *frequency matrices* of the modified equation (10).

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,*

$$\|w(\cdot, t)\|_{L^2(\Omega)} \leq C \|w_0(\cdot)\|_{L^2(\Omega)} \quad (23)$$

*holds for a positive constant  $C$  if*

$$\text{Real}(\mu_{k,i}) < 0$$

*for all eigenvalues  $\mu_{k,i}$  of  $\mathcal{A}_k$  with  $k \in \mathbb{Z}^{\neq 0}$ .*

*Remark 4* One might argue that Corollary 1 is not needed in the sense that for every matrix  $B$  with  $B$  positive definite, there holds (23). However, this is 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  $x^T B x < 0$  for, e.g.,  $x := (1, 10)^T$ ), however, the eigenvalues of  $\mathcal{A}_k$  have negative real part, and consequently, the complete system (19) is stable. (A tedious computation reveals that the eigenvalues of  $\mathcal{A}_k$  are  $2\pi k ((\pm\sqrt{17} - 5)\pi k - i)$ .)

As already pointed out in the introduction, we have the following important remark concerning commutative matrices:

*Remark 5* The real part of the eigenvalues of  $\mathcal{A}_k$  is *not* affected by the terms stemming from  $A$  if matrices  $A$  and  $B$  can be simultaneously diagonalized. This is the motivation for introducing so called *characteristic splittings* in the following section.

## 4 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  $\varepsilon$  *without* any additional stabilization terms. The splitting relies on a characteristic decomposition of the matrix  $A$ , i.e.,  $A$  can be decomposed into

$$A = Q\Lambda Q^{-1} \quad (24)$$

for an invertible  $Q$  and  $\Lambda := \text{diag}(\lambda_1, \dots, \lambda_d)$ . The idea of the characteristic splitting is to split the matrix  $\Lambda$  into stiff and nonstiff parts. We make this more precise in the following definition:

**Definition 3** Let  $A$  be decomposed as in (24), and let  $A$  be split into

$$A = \hat{A} + \tilde{A}$$

in such a way that  $\hat{A}$  and  $\tilde{A}$  are diagonal matrices and define an admissible splitting of  $A$  in the sense of Definition 1. (Consequently, the entries of  $\hat{A}$  can be bounded independently of  $\varepsilon$ .) Subsequently, the *characteristic splitting* is defined by

$$\hat{A} = Q\hat{\Lambda}Q^{-1} \quad \text{and} \quad \tilde{A} = Q\tilde{\Lambda}Q^{-1}. \quad (25)$$

Obviously, the splitting of  $A$  is admissible in the sense of Definition 1.

Let us make the following remark about characteristic splittings:

- Remark 6* – As the system (1) is hyperbolic for all  $\varepsilon > 0$ , one can always obtain an admissible characteristic splitting by choosing  $\hat{A}$  as  $A|_{\varepsilon=1}$ , and  $\tilde{A} := A - \hat{A}$ . Obviously, the eigenvalues of both  $\hat{A}$  and  $\tilde{A}$  are real, and those of  $\hat{A}$  are trivially independent of  $\varepsilon$ . Such a decomposition would lead to a fully explicit scheme for  $\varepsilon = 1$ , which is desirable as for nonstiff equations, those schemes usually are less diffusive than implicit ones.
- Note that  $Q$  still depends on  $\varepsilon$ , and so, even if  $\hat{A}$  is independent of  $\varepsilon$ ,  $\tilde{A}$  generally is not (but its eigenvalues are).
  - As one can see from Section 5 and especially Section 6, a crucial part in our analysis is the fact that  $\hat{A}$  and  $\tilde{A}$  commute, i.e.,  $\tilde{A}\hat{A} = \hat{A}\tilde{A}$ . In this way, the 'bad' modes that can potentially destroy uniform stability are ruled out.

In the sequel, we apply this concept to a prototype matrix  $A$ , given by

$$A = \begin{pmatrix} a & 1 & 0 \\ \frac{1}{\varepsilon^2} & a & \frac{1}{\varepsilon^2} \\ 0 & 1 & a \end{pmatrix}. \quad (26)$$

Its eigenvalues are

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

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

$$\begin{aligned} \hat{A} &:= \text{diag} \left( a - \sqrt{2}, a, a + \sqrt{2} \right), \\ \tilde{A} &:= \text{diag} \left( -\frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}, 0, \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon} \right). \end{aligned}$$

Consequently, we can derive matrices  $\hat{A}$  and  $\tilde{A}$  via (25) as

$$\hat{A} = \begin{pmatrix} a & \varepsilon & 0 \\ \frac{1}{\varepsilon} & a & \frac{1}{\varepsilon} \\ 0 & \varepsilon & a \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} 0 & 1 - \varepsilon & 0 \\ \frac{1 - \varepsilon}{\varepsilon^2} & 0 & \frac{1 - \varepsilon}{\varepsilon^2} \\ 0 & 1 - \varepsilon & 0 \end{pmatrix}.$$

The focus of this paper is on uniform stability as  $\varepsilon \rightarrow 0$ , 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 (4), or the instability due to condition (5). In the following (cf. Theorem 2), we derive upper bounds on the nonstiff CFL number that assure stability (in a sense to be made more precise) of IMEX scheme (7) for a characteristic splitting.

## 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 (7) is stable. We start with the general case, and subsequently consider the prototype equation.

### 5.1 General case

We consider the characteristic splitting (25) in the light of Corollary 1. For a generic splitting with *commuting* matrices  $\hat{A}$  and  $\tilde{A}$ , the frequency matrix  $\mathcal{A}_k$  (see (22) and (17)) can be written as

$$\mathcal{A}_k = -i2\pi k A - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - (\hat{A} - \tilde{A})(\hat{A} + \tilde{A}) \right) \quad (28)$$

$$= -i2\pi k A - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - \hat{A}^2 + \tilde{A}^2 \right). \quad (29)$$

Note that  $\mathcal{A}_0 = 0$ , since constant (in space) solutions of the modified equations are constant in time also. Therefore we need to analyze only the case  $k \neq 0$ . As we rely on a characteristic splitting, all the matrices occurring in (29) can be written as  $Q\Sigma Q^{-1}$  for some diagonal matrix  $\Sigma$ . Thus, it is easy to see that the real part  $\mu_{k,i}$  of the eigenvalues of  $\mathcal{A}_k$  is given by

$$\text{Real}(\mu_{k,i}) = 2\pi^2 k^2 \Delta t \left( -\frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} + \hat{\lambda}_i^2 - \tilde{\lambda}_i^2 \right)$$

where  $\hat{\lambda}_i$  and  $\tilde{\lambda}_i$  are eigenvalues to  $\hat{A}$  and  $\tilde{A}$ , respectively. Claiming that  $\text{Real}(\mu_{k,i})$  is negative leads to

$$\frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} > \hat{\lambda}_i^2 - \tilde{\lambda}_i^2. \quad (30)$$

This leads to a time step restriction that only depends on the *explicit* part. We summarize this in the following lemma:

**Lemma 3** *Let  $\frac{\Delta t}{\Delta x}$  be restricted by*

$$\frac{\Delta t}{\Delta x} < \frac{\widehat{\alpha} + \widetilde{\alpha}}{\max_i \widehat{\lambda}_i^2} \quad (31)$$

*Then, (30) and thus Corollary 1 hold.*

This is a good result, as  $\widehat{\lambda}_i^2$  can be bounded independently of  $\varepsilon$ , and therefore, (31) is also independent of  $\varepsilon$ . We summarize this in the following theorem.

**Theorem 2** *The characteristic splitting as introduced in (25) is such that  $\text{Real}(\mu_{k,i}) < 0$  holds for all  $k \in \mathbb{Z}^{\neq 0}$ ,  $\mu_{k,i}$  eigenvalue to  $\mathcal{A}_k$ , with a restriction on the time step size that is independent of  $\varepsilon$ .*

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

## 5.2 Characteristic Splitting of prototype equation

In this section, we consider the prototype matrix from Section 4, as it allows for easy and explicit computations. The non-dimensionalized advective CFL number corresponding to  $A$  is denoted by

$$\widehat{\nu} := \frac{a\Delta t}{\Delta x}. \quad (32)$$

Note that using  $\widehat{\nu}$ , (31) reads

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

In the sequel, we give stronger bounds on the allowable time step size. Given  $k \in \mathbb{Z}^{\neq 0}$ , we consider the frequency matrix  $\mathcal{A}_k$  for the characteristic splitting introduced in (25). 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  $\{\mu_{k,0}, \mu_{k,\pm}\}$  of  $\mathcal{A}_k$  are given by*

$$\text{Real}(\mu_{k,0}) = -2\pi^2 k^2 \Delta x (\widehat{\alpha} + \widetilde{\alpha}) + 2\pi^2 k^2 \Delta t a^2 \quad (33a)$$

$$\begin{aligned} \text{Real}(\mu_{k,\pm}) &= \frac{-4\pi^2 k^2 \Delta t}{\varepsilon^2} + \frac{8\Delta t k^2 \pi^2}{\varepsilon} \\ &\quad + 2\pi^2 k^2 a^2 \Delta t - 2\pi^2 k^2 \Delta x (\widehat{\alpha} + \widetilde{\alpha}) \pm 4\sqrt{2} a \pi^2 k^2 \Delta t. \end{aligned} \quad (33b)$$

*Proof* With the notation introduced in (24) and (25), see also (29), there holds

$$\begin{aligned}\mathcal{A}_k &= -i2\pi k Q \Lambda Q^{-1} - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - Q(\hat{\Lambda} - \tilde{\Lambda}) \Lambda Q^{-1} \right) \\ &= Q \left( -i2\pi k \Lambda - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} \text{Id} - (\hat{\Lambda} - \tilde{\Lambda}) \Lambda \right) \right) Q^{-1} \\ &= Q \text{diag}(\sigma) Q^{-1},\end{aligned}$$

where the vector  $\sigma$  is given by

$$\sigma := \begin{pmatrix} -i2\pi k(a - \frac{\sqrt{2}}{\varepsilon}) - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} + \frac{2}{\varepsilon^2} - \frac{4}{\varepsilon} + 2\sqrt{2}a - a^2 \right) \\ -i2\pi k a - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} - a^2 \right) \\ -i2\pi k(a + \frac{\sqrt{2}}{\varepsilon}) - 2\pi^2 k^2 \Delta t \left( \frac{(\hat{\alpha} + \tilde{\alpha}) \Delta x}{\Delta t} + \frac{2}{\varepsilon^2} - \frac{4}{\varepsilon} - 2\sqrt{2}a - a^2 \right) \end{pmatrix}$$

Thus, one can conclude that the eigenvalues of  $\mathcal{A}_k$  are given in  $\sigma$ . Sorting the eigenvalues conveniently, starting with the one in the middle, one can conclude that their Real parts are given by formulae (33a) and (33b).  $\square$

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

**Lemma 5**  $\text{Real}(\mu_{k,0}) < 0$  under the CFL condition

$$\hat{\nu} < \nu_1 := \frac{\hat{\alpha} + \tilde{\alpha}}{a}. \quad (34)$$

*Remark 7* Condition (34) is a condition for the nonstiff CFL number  $\hat{\nu}$  from (32). It depends on both  $\hat{\alpha}$  and  $\tilde{\alpha}$ . The coefficient  $\hat{\alpha}$  encodes the upwind viscosity of the explicit numerical flux (8), 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 (9), and limiting choices are either  $\tilde{\alpha} = \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}$  (the largest eigenvalue of the nonstiff matrix) or  $\tilde{\alpha} = 0$ . In both cases,  $\hat{\alpha} + \tilde{\alpha} \geq a + \sqrt{2}$ , which gives the sufficient stability condition

$$\hat{\nu} < \frac{a + \sqrt{2}}{a}.$$

This is independent of  $\varepsilon$ .

Now we discuss  $\text{Real}(\mu_{k,\pm})$ . Obviously, for  $\Delta t$  fixed and  $\varepsilon \rightarrow 0$ ,  $\text{Real}(\mu_{k,\pm}) < 0$ , which directly yields the following Proposition:

**Proposition 1** *Let  $\Delta t$  be fixed. Then, there exists an  $\varepsilon_0 > 0$ , such that for all  $\varepsilon < \varepsilon_0$  and all  $k \in \mathbb{Z}^{\neq 0}$ ,  $\text{Real}(\mu_{k,\pm}) < 0$ .*

However, this is not the full asymptotics. We therefore change the point of view: Given a *fixed*  $\varepsilon$ , for which  $\Delta t$  is  $\text{Real}(\mu_{k,\pm}) < 0$ ? The following lemma provides the crucial estimate:

**Lemma 6** *We define*

$$\varphi(a) := \frac{\sqrt{2}}{a + 2\sqrt{2}}. \quad (35)$$

*Now, consider two cases as follows:*

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

$$\hat{\nu} \leq \frac{(\hat{\alpha} + \tilde{\alpha})a}{(a + \sqrt{2})^2}.$$

*Proof* For  $\text{Real}(\mu_{k,\pm})$  to be negative, it suffices to show that

$$0 > \frac{-4\Delta t}{\varepsilon^2} + \frac{8\Delta t}{\varepsilon} + 2a^2\Delta t - 2\Delta x(\hat{\alpha} + \tilde{\alpha}) + 4\sqrt{2}a\Delta t.$$

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

$$\begin{aligned} 0 &> \frac{-4}{\varepsilon^2} + \frac{8}{\varepsilon} + 2a^2 - \frac{2(\hat{\alpha} + \tilde{\alpha})a}{\hat{\nu}} + 4\sqrt{2}a \\ \Leftrightarrow \frac{(\hat{\alpha} + \tilde{\alpha})a}{\hat{\nu}} &> \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a. \end{aligned} \quad (36)$$

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

$$\begin{aligned} 0 &\geq \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a \\ \Leftrightarrow 0 &\geq -2 + 4\varepsilon + \varepsilon^2 (a^2 + 2\sqrt{2}a) \\ \Leftrightarrow 0 &\leq \varepsilon \leq \varphi(a). \end{aligned}$$

This proves the first claim that for all  $\varepsilon \leq \varphi(a)$ , both  $\text{Real}(\mu_{k,\pm})$  are negative.

Now let  $\varphi(a) < \varepsilon$ . In this case, the right-hand side of (36) is positive, and therefore one has a restriction on  $\hat{\nu}$ . One can compute

$$\begin{aligned} \frac{(\hat{\alpha} + \tilde{\alpha})a}{\hat{\nu}} &> \frac{-2}{\varepsilon^2} + \frac{4}{\varepsilon} + a^2 + 2\sqrt{2}a \\ \Leftrightarrow \frac{\hat{\nu}}{(\hat{\alpha} + \tilde{\alpha})a} &< \frac{1}{\frac{-2+4\varepsilon}{\varepsilon^2} + a^2 + 2\sqrt{2}a}. \end{aligned} \quad (37)$$

Note that for any  $0 \leq \varepsilon \leq 1$ , there holds  $\frac{-2+4\varepsilon}{\varepsilon^2} \leq 2$ . Consequently, (37) is fulfilled if

$$\frac{\hat{\nu}}{(\hat{\alpha} + \tilde{\alpha})a} \leq \frac{1}{2 + a^2 + 2\sqrt{2}a} = \frac{1}{(a + \sqrt{2})^2}$$

This proves the lemma.  $\square$

With Lemmas 5 and 6 we obtain the following

**Proposition 2** *Recall the definition of  $\nu_1$  from (34). For  $k \in \mathbb{Z}^{\neq 0}$ ,  $\text{Real}(\mu_{k,0}) < 0$  and  $\text{Real}(\mu_{k,\pm}) < 0$  if*

$$\widehat{\nu} < \nu_2 := \nu_1 \psi(\varepsilon, a) \quad (38)$$

with

$$\psi(\varepsilon, a) := \begin{cases} 1, & \varepsilon \leq \varphi(a) \\ \left(\frac{a}{a+\sqrt{2}}\right)^2 & \varepsilon > \varphi(a). \end{cases} \quad (39)$$

From the previous considerations, we can conclude:

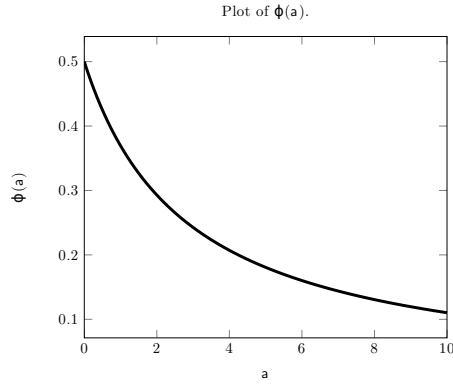
**Corollary 2** *We choose  $\widehat{\alpha} = a + \sqrt{2}$ . Then, there holds  $\text{Real}(\mu_{k,i}) < 0$  for all  $k \in \mathbb{Z}^{\neq 0}$ ,  $\mu_{k,i}$  eigenvalue to  $\mathcal{A}_k$ , if*

$$\frac{(a + \sqrt{2})\Delta t}{\Delta x} < 1. \quad (40)$$

*Proof* Consider expression (38) and plug in the definition of both  $\psi$  from (39) and  $\nu_1$  from (34). We consider case  $\varepsilon > \varphi(a)$  first, and obtain

$$\left(\frac{\widehat{\alpha} + \widetilde{\alpha}}{a}\right) \left(\frac{a}{a + \sqrt{2}}\right)^2 \geq \frac{(a + \sqrt{2})a}{(a + \sqrt{2})^2} \geq \frac{a}{a + \sqrt{2}}.$$

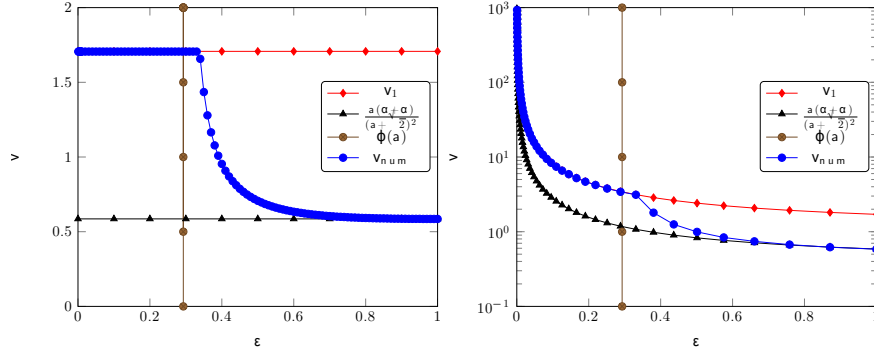
Ergo,  $\widehat{\nu} < \frac{a}{a+\sqrt{2}}$  is sufficient for (38), which implies (40). Similarly, for  $\varepsilon \leq \varphi(a)$ , one can easily show that  $\widehat{\nu} < \nu_1$  is fulfilled given that (40) holds.  $\square$



**Fig. 1** Plot of function  $\varphi$  from (35).

*Remark 8* – The function  $\varphi(a) := \frac{\sqrt{2}}{a+2\sqrt{2}}$ , see (35), has been plotted in Figure 1. Consider the stability constraint (38), together with the definition of  $\psi(\varepsilon, a)$  in (39). From Figure 1, one can tell that  $\varepsilon \leq \varphi(a)$  in (39) is met for a sufficiently small  $\varepsilon$ . This then again implies that for ‘small’  $\varepsilon$ , one has stability that only depends on the convective CFL number  $\nu_1$ , see (34), as  $\psi(\varepsilon) = 1$  for this case. This is an impressive result in the sense that the only stability restriction for small  $\varepsilon$  depends on the slow waves.

- In Figure 2, we plotted the numerically determined maximum allowable  $\hat{\nu}_{num}$  values such that the real parts of the eigenvalues of  $\mathcal{A}_k$  are negative. For the particular computations, we choose  $a = 2$ ,  $\Delta x = 10^{-2}$ ,  $\Delta t = \frac{\hat{\nu}_{num} \Delta x}{a}$ ,  $\hat{\alpha} = 2 + \sqrt{2}$ ,  $\tilde{\alpha} = 0$  or  $\tilde{\alpha} = \frac{\sqrt{2}(1-\varepsilon)}{\varepsilon}$  and determine the maximum  $\hat{\nu}_{num}$ , such that  $\text{Real}(\mu_{k,0}) < 0$  and  $\text{Real}(\mu_{k,\pm}) < 0$ . One can infer from this figure that the stability restriction one has to impose on the ratio  $\frac{\Delta t}{\Delta x}$  can be made independent of  $\varepsilon$ . Note that if there is no known explicit formula for the eigenvalues of  $\mathcal{A}_k$ , one could still investigate linearized splittings of, e.g., the Euler equations by simply computing all  $\mathcal{A}_k$  up to a given value of  $k$ , getting a first glimpse of possible (in)stabilities in the splitting.



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

## 6 On The Non-Uniform Stability Of Some Splittings

In this section, we consider a splitting that does not induce a scheme that is uniformly stable. This means that there is no bound on  $\hat{\nu}$ , independently of  $\varepsilon$ , such that  $\text{Real}(\mu_{k,i}) < 0$  holds for all  $k \in \mathbb{Z}^{\neq 0}$  and all  $\varepsilon > 0$ .

In particular, we consider the (non-characteristic) splitting of matrix  $A$  from (26) into

$$\hat{A} = \begin{pmatrix} a & 1-\varepsilon & 0 \\ 1 & a & 1 \\ 0 & 1-\varepsilon & a \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} 0 & \varepsilon & 0 \\ \frac{1-\varepsilon^2}{\varepsilon^2} & 0 & \frac{1-\varepsilon^2}{\varepsilon^2} \\ 0 & \varepsilon & 0 \end{pmatrix}.$$



The eigenvalues of the splitting matrices are

$$\begin{aligned}\widehat{\lambda}_1 &= a, & \widehat{\lambda}_{2,3} &= a \pm \sqrt{2 - 2\varepsilon} \\ \widetilde{\lambda}_1 &= 0, & \widetilde{\lambda}_{2,3} &= \pm \frac{\sqrt{2\varepsilon(1 - \varepsilon^2)}}{\varepsilon}.\end{aligned}$$

Obviously, this only yields a hyperbolic splitting in the sense of Definition 1 if we restrict ourselves to  $\varepsilon < 1$ . (We are only interested in the case  $\varepsilon \rightarrow 0$ , so this can be done without loss of generality, as it could also be circumvented by a reparametrization of  $\varepsilon$ .) For  $\varepsilon < 1$ , the splitting is admissible.

The frequency matrix  $\mathcal{A}_k$  in this case has eigenvalues which can only be computed via an extremely tedious calculation, or with the aid of Maple. Expanded in terms of the asymptotic sequence  $\{\varepsilon^{-2}, \varepsilon^{-1}, \dots\}$ , these eigenvalues are given by

$$\begin{aligned}\mu_{k,0} &= 2\pi^2 k^2 \Delta t a^2 - 2\pi^2 k^2 (\widehat{\alpha} + \widetilde{\alpha}) \Delta x - 2\pi k i a + O(\varepsilon), \\ \mu_{k,+} &= \frac{4\pi^2 k^2 \Delta t}{\varepsilon^2} - \frac{8\pi^2 k^2 \Delta t}{\varepsilon} + O(1), & \mu_{k,-} &= -\frac{4\pi^2 k^2 \Delta t}{\varepsilon^2} + O(1).\end{aligned}\quad (41)$$

Note that  $\text{Real}(\mu_{k,+}) < 0$  for  $\varepsilon \rightarrow 0$  can only hold if  $\Delta t = O(\varepsilon)$  (and so the asymptotic expansion is not valid anymore, because  $O(1)$  refers to  $O(1)$  with respect to  $\varepsilon$ , not with respect to  $\Delta t$ ). Consequently, a CFL condition independently of  $\varepsilon$  can *not* hold, although we treat the 'fast' parts implicitly.

*Remark 9* – This result is in contrast to common belief that coupling two schemes that are individually stable, does indeed yield a stable scheme.

- In particular, recall the two forms of the frequency matrix  $\mathcal{A}_k$  in (28) and (29). They differ by the commutator

$$-2\pi^2 k^2 \Delta t (\widetilde{A}\widehat{A} - \widehat{A}\widetilde{A}),$$

whose eigenvalues are

$$0, \pm(4\pi^2 k^2 \Delta t) \frac{1 - \varepsilon - \varepsilon^2}{\varepsilon^2} = \pm \frac{4\pi^2 k^2 \Delta t}{\varepsilon^2} + O(\varepsilon^{-1}),$$

and this is precisely the leading order term of  $\mu_{k,\pm}$  in (41). This seems to indicate the importance of the commutator, and the difficulty to control its contribution to the frequency matrix.

- Our result is a consequence of the fact that for two matrices  $A$  and  $B$ , there is no bound on the eigenvalues of  $A \cdot B$  in terms of products of eigenvalues of  $A$  and  $B$ . In our example, the eigenvalues of the commutator are asymptotically larger than the product of those of  $\widehat{A}$  and  $\widetilde{A}$ . More precisely,

$$\frac{4\pi^2 k^2 \Delta t}{\varepsilon^2} \gg |\widehat{\lambda}_{2,3}| |\widetilde{\lambda}_{2,3}| = \frac{\sqrt{2}(a + \sqrt{2})}{\sqrt{\varepsilon}} + O(\sqrt{\varepsilon}).$$

- Once more, the characteristic splitting removes the commutator.

## 7 On The Non-Uniform Stability Of Splittings For The Linearized Euler Equations

In this section, we apply the theory developed earlier to the linearized Euler equations.

### 7.1 Problem statement and analysis

From now on, we consider  $\gamma$  to be the fixed constant  $\gamma = 1.4$ . Considering a particular simple state

$$v_0 := (\rho_0, \rho_0 u_0, E_0) := (1, 1, 1), \quad (42)$$

and setting  $A := f'(v_0)$ , where  $f$  is defined in (3), we obtain the linearized system (1) with matrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{3}{2} + \frac{1}{2}\gamma & 3 - \gamma & \frac{\gamma-1}{\varepsilon^2} \\ \gamma\varepsilon^2 - \varepsilon^2 - \gamma & \gamma - \frac{3}{2}\gamma\varepsilon^2 + \frac{3}{2}\varepsilon^2 & \gamma \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 0 & 5 & 0 \\ -4 & 8 & \frac{2}{\varepsilon^2} \\ 2\varepsilon^2 - 7 & 7 - 3\varepsilon^2 & 7 \end{pmatrix}.$$

Its eigenvalues are

$$\lambda_1 = 1 \quad (43)$$

$$\lambda_{\pm} = 1 \pm \frac{\sqrt{\gamma(\gamma-1)(1-\frac{\varepsilon^2}{2})}}{\varepsilon} = 1 \pm \frac{\sqrt{0.56(1-\frac{\varepsilon^2}{2})}}{\varepsilon} \quad (44)$$

and consequently, the associated system of conservation laws fits very nicely into our framework with two fast waves and one slow convective wave.

We consider a splitting taken from literature [27], which is actually a modification of Klein's splitting [22]. On the nonlinear level, it is given by a splitting of the flux function  $f(v)$  into the sum of

$$\widehat{f}(v) := (\rho u, \rho u^2 + p, u(E + \Pi))^T, \quad (45a)$$

$$\widetilde{f}(v) := \left(0, \frac{1-\varepsilon^2}{\varepsilon^2}p, u(p - \Pi)\right)^T. \quad (45b)$$

$\Pi$  is an auxiliary pressure function, and it is defined by

$$\Pi(x, t) := \varepsilon^2 p(x, t) + (1 - \varepsilon^2)\bar{p} \quad (46)$$

for a constant value (with respect to space) of  $\bar{p}$ . We cannot completely mimic the nonlinear behavior, as this value is often chosen as the infimum of the pressure over the spatial domain. However, we can set it to the constant value of  $\bar{p} = \frac{1}{5}$ , which is the pressure for  $\varepsilon = 1$ , and is the infimum for all  $0 \leq$

$\varepsilon \leq 1$ . Using this (arguably crude) choice, it is straightforward to linearize the splittings, and one obtains the non-stiff matrix

$$\hat{A} = \frac{1}{5} \begin{pmatrix} 0 & 5 & 0 \\ -5 + \varepsilon^2 & 10 - 2\varepsilon^2 & 2 \\ -6 - \varepsilon^2 + 2\varepsilon^4 & 6 + \varepsilon^2 - 3\varepsilon^4 & 5 + 2\varepsilon^2 \end{pmatrix} \quad (47)$$

with eigenvalues

$$\hat{\lambda}_1 = 1 \quad (48)$$

$$\hat{\lambda}_{2,3} = 1 \pm \frac{1}{5} \sqrt{12 - 3\varepsilon^2 - 2\varepsilon^4} \quad (49)$$

and the corresponding stiff matrix

$$\tilde{A} = \frac{1}{5} \begin{pmatrix} 0 & 0 & 0 \\ 1 - \varepsilon^2 & -2 + 2\varepsilon^2 & -\frac{2(\varepsilon^2 - 1)}{\varepsilon^2} \\ -1 + 3\varepsilon^2 - 2\varepsilon^4 & 1 - 4\varepsilon^2 + 3\varepsilon^4 & 2 - 2\varepsilon^2 \end{pmatrix} \quad (50)$$

with eigenvalues

$$\tilde{\lambda}_1 = 0 \quad (51)$$

$$\tilde{\lambda}_{2,3} = \pm \frac{(\varepsilon^2 - 1)\sqrt{2 - 2\varepsilon^2}}{5\varepsilon}. \quad (52)$$

Using Maple, we can easily evaluate the eigenvalues  $\mu$  of the frequency matrix  $\mathcal{A}_k$  and approximately writing them as

$$\mu_{k,0} = O(1) \quad (53)$$

$$\mu_{\pm} \approx \frac{(-1.579136704 \pm 9.474820224)k\Delta t}{\varepsilon^2} + O(\varepsilon^{-1}). \quad (54)$$

The same results as in Section 6 holds, and  $\text{Real}(\mu_{k,+}) < 0$  can only hold for  $\Delta t = O(\varepsilon)$ .

## 7.2 Numerical Results

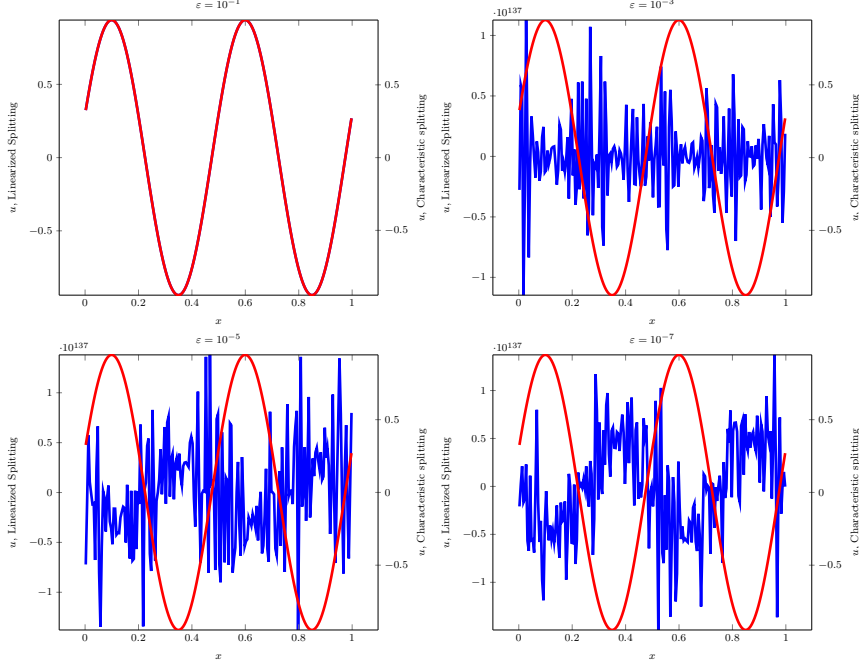
In this section, we substantiate the results from the previous subsection with suitable numerical experiments. The setup is as before on domain  $\Omega = [0, 1]$ , and we consider the linearized splitting defined by the matrices  $\hat{A}$  and  $\tilde{A}$  in (47) and (50), respectively. In addition, we consider a characteristic splitting, where  $\hat{A}$  is defined as the diagonal matrix with eigenvalues corresponding to  $\varepsilon = 1$ . Initial data are given in the *characteristic* variables  $w$  as

$$w(x, 0) = (\cos(4\pi x), 0, 0)^T \quad (55)$$

where the first component corresponds to the 'slow' eigenvalue. Note that with these initial conditions, it is guaranteed that there is a limit solution for  $\varepsilon \rightarrow 0$ .

$\hat{\alpha}$	$\hat{\alpha}$	$\Delta x$	$\Delta t$	$T_{end}$
Maximum absolute eigenvalues of $\hat{A}$	0	1/200	$10^{-1} \Delta x$	0.1

**Table 1** Parameters used for the computations in Figure 3.



**Fig. 3** Comparison of classical versus characteristic splitting. Blue: Results based on linearized splitting. Red: Results based on characteristic splitting. Note the different scales in the plot.

Numerical results are shown in Figure 3. Those results have been computed with the set of parameters as given in Table 1. Note that the choice of  $\Delta t$  corresponds to a non-stiff cfl number of  $\frac{1.53}{10} = 0.153$  for the characteristic splitting, and approximately  $\frac{1.7}{10} = 0.17$  for the linearized splitting.

It is clearly visible that the linearized splitting is unstable, at least with the uniform choice of  $\Delta t$  independent of  $\varepsilon$ , while the characteristic one is not. (Note: The left  $y$  axis corresponds to the linearized splitting, and the right one to the characteristic one.)

## 8 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. The eigenvalue analysis reveals the

subtle interplay of terms stemming from the discretization of both advection and diffusion, and an additional term stemming from truncation errors in time.

Our analysis, in contrast to common belief, shows that the nonstiff CFL number is usually *not* enough to ensure stability of an IMEX scheme. Indeed, for a splitting introduced earlier, the analysis shows that there is a time step restriction of (at least) order  $\varepsilon$  to ensure stability, so the resulting algorithm is not asymptotically stable.

To circumvent this problem, we introduced a new way of obtaining suitable splittings via characteristic decomposition of the flux Jacobian. Those splittings are stable under a constraint on the nonstiff CFL number that is *independent* of  $\varepsilon$ . 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. Another challenge is the treatment of multiple dimensions, as the flux-Jacobians usually do not commute. A suitable extension is subject to current research.

The results presented in this paper concern only the issue of asymptotic stability. In future work we will extend this study and include the quality of the approximation of low- and higher-order IMEX schemes for 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., Ruuth, S., Spiteri, R.: 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. Choi, Y.H., Merkle, C.: The application of preconditioning in viscous flows. *Journal of Computational Physics* **105**(2), 207 – 223 (1993)
5. Chorin, A.: The numerical solution of the Navier-Stokes equations for an incompressible fluid. *Bulletin of the American Mathematical Society* **73**, 928–931 (1967)
6. Colella, P., Pao, K.: A projection method for low speed flows. *Journal of Computational Physics* **149**(2), 245–269 (1999)
7. 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)
8. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen* **100**(1), 32–74 (1928)
9. Crouzeix, M.: Une méthode multipas implicite-explicite pour l’approximation des équations d’évolution paraboliques. *Numerische Mathematik* **35**(3), 257–276 (1980)
10. Degond, P., Łozinski, 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)
11. 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)

12. Dellacherie, S.: Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *Journal of Computational Physics* **229**(4), 978 – 1016 (2010)
13. Godlewski, E., Raviart, P.A.: *Hyperbolic Systems of Conservation Laws*. Ellipses Paris (1991)
14. Guillard, H., Murrone, A.: On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes. *Computers and Fluids* **33**(4), 655 – 675 (2004)
15. Guillard, H., Viozat, C.: On the behaviour of upwind schemes in the low Mach number limit. *Computers and Fluids* **28**(1), 63–86 (1999)
16. 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)
17. Hairer, E., Wanner, G.: *Solving Ordinary Differential Equations II*. Springer Series in Computational Mathematics (1991)
18. Jin, S.: Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing* **21**, 441–454 (1999)
19. Jin, S.: Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: A review. *Rivista di Matematica della Università di Parma* **3**, 177–216 (2012)
20. Jin, S., Pareschi, L., Toscani, G.: Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM Journal on Numerical Analysis* **35**, 2405–2439 (1998)
21. Klainerman, S., Majda, A.: Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Communications on Pure and Applied Mathematics* **34**, 481–524 (1981)
22. 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)
23. Klein, R., Botta, N., Schneider, T., Munz, C., Roller, S., Meister, A., Hoffmann, L., Sonar, T.: Asymptotic adaptive methods for multi-scale problems in fluid mechanics. *Journal of Engineering Mathematics* **39**(1), 261–343 (2001)
24. Kröner, D.: *Numerical Schemes for Conservation Laws*. Wiley Teubner (1997)
25. 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)
26. Murrone, A., Guillard, H.: Behavior of upwind scheme in the low Mach number limit: III. Preconditioned dissipation for a five equation two phase model. *Computers and Fluids* **37**(10), 1209 – 1224 (2008)
27. Noelle, S., Bispin, 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)
28. Richtmyer, R., Morton, K.: *Difference methods for initial-value problems*. Krieger Publishing Company (1994)
29. 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)
30. Schütz, J.: An asymptotic preserving method for linear systems of balance laws based on Galerkin’s method. *Journal of Scientific Computing* **60**, 438–456 (2014). DOI 10.1007/s10915-013-9801-1
31. Strang, G.: Accurate partial difference methods. *Numerische Mathematik* **6**, 37–46 (1964)
32. Turkel, E.: Preconditioned methods for solving the incompressible and low speed compressible equations. *Journal of Computational Physics* **72**(2), 277 – 298 (1987)
33. 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)