Invariant-Domain Preserving High-Order Time Stepping: II. IMEX Schemes

Alexandre Ern, Jean-Luc Guermond

Invariant-domain preserving high-order time stepping: II. IMEX schemes*

Alexandre Ern† and Jean-Luc Guermond‡

Draft version June 23, 2022

Abstract

We consider high-order discretizations of a Cauchy problem where the evolution operator comprises a hyperbolic part and a parabolic part with diffusion and stiff relaxation terms. We propose a technique that makes every implicit-explicit (IMEX) time stepping scheme invariant-domain preserving and mass conservative. Following the ideas introduced in Part I on explicit Runge–Kutta schemes, the IMEX scheme is written in incremental form. At each stage, we first combine a low-order and a high-order hyperbolic update using a limiting operator, then we combine a low-order and a high-order parabolic update using another limiting operator. The proposed technique, which is agnostic to the space discretization, allows to optimize the time step restrictions induced by the hyperbolic sub-step. To illustrate the proposed methodology, we derive four novel IMEX methods with optimal efficiency. All the implicit schemes are singly diagonal. One of them is A-stable and the other three are L-stable. The novel IMEX schemes are evaluated numerically on a stiff ODE system and a scalar nonlinear conservation equation.

Keywords. Time integration, implicit-Explicit time integration methods, conservation equations, hyperbolic systems, invariant-domains, high-order method.

MSC. 35L65, 65M60, 65M12, 65N30

1 Introduction

This work is the second part of a project started in [10] whose objective is to develop Runge–Kutta time stepping schemes that are invariant-domain preserving (IDP) and conservative. The scope of the present work lies in the approximation of the following Cauchy problem posed on the space domain $D \subset \mathbb{R}^d$ and the time interval $J := (0, T)$ with $T > 0$:

$$\partial_t \boldsymbol{u} + \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{g}(\boldsymbol{u}, \nabla \boldsymbol{u}) = \boldsymbol{0} \text{ in } D \times J, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0 \text{ in } D, \tag{1}$$

supplemented with appropriate boundary conditions. The dependent variable $\boldsymbol{u}$ takes values in $\mathbb{R}^m$ with $m \geq 1$. The operator $\boldsymbol{f} : \mathcal{A} \to \mathbb{R}^m$ represents the hyperbolic part of the problem, and

†CERMICS, Ecole des Ponts, 77455 Marne-la-Vallee Cedex 2, France and INRIA Paris, 75589 Paris, France

‡Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA.

*This material is based upon work supported in part by the National Science Foundation grant DMS2110868, the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397, the Army Research Office, under grant number W911NF-19-1-0431, and the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contracts B640889. The support of INRIA through the International Chair program is acknowledged.

the operator $\boldsymbol{g} : \mathcal{A} \times \mathbb{R}^{m\times d} \to \mathbb{R}^m$ represents the parabolic part, typically associated with diffusion and (stiff) relaxation processes. Here, $\mathcal{A}$ is the domain of $\mathbb{f}$ and $\mathcal{A} \times \mathbb{R}^{m\times d}$ is the domain of $\mathbb{g}$. In the applications we have in mind, these operators have the following structure:

$$\boldsymbol{f}(\boldsymbol{u}) = \nabla\cdot\mathbb{f}(\boldsymbol{u}), \qquad \boldsymbol{g}(\boldsymbol{u}, \nabla\boldsymbol{u}) = \nabla\cdot\mathbb{d}(\boldsymbol{u}, \nabla\boldsymbol{u}) + \boldsymbol{r}(\boldsymbol{u}), \tag{2}$$

with the hyperbolic flux $\mathbb{f} : \mathcal{A} \to \mathbb{R}^{m\times d}$, the diffusive flux $\mathbb{d} : \mathcal{A} \times \mathbb{R}^{m\times d} \to \mathbb{R}^{m\times d}$, and the relaxation operator $\boldsymbol{r} : \mathcal{A} \to \mathbb{R}^m$.

As it is out of the scope of this paper to discuss the existence and uniqueness of solutions to (1), we assume that this problem admits a reasonable class of solutions. We also assume that the set $\mathcal{A} \subset \mathbb{R}^m$ is an invariant domain for this solution class. This means that if $\boldsymbol{u}_0(\boldsymbol{x}) \in \mathcal{A}$ for a.e. $\boldsymbol{x}$ in D (and up to perturbations resulting from boundary conditions which go beyond the scope of this paper), then any admissible solution to (1) takes values in $\mathcal{A}$ at a.e. $\boldsymbol{x}$ in D at a.e. time $t \in [0, T]$. The set $\mathcal{A}$ may depend on $\boldsymbol{u}_0$. A simple example is the scalar convection-diffusion equation (i.e., $m = 1$), in which case the interval $\mathcal{A} := [\operatorname{ess\,inf}_{x\in D} u_0(x), \operatorname{ess\,sup}_{x\in D} u_0(x)] \subset \mathbb{R}$ is an invariant domain. Two more elaborate examples are the compressible Euler equations and the compressible Navier–Stokes equations. For these equations, the conserved variable $\boldsymbol{u}$ takes values in $\mathbb{R}^{d+2}$ and its components are the density, the momentum, and the total mechanical energy (i.e., $m = d + 2$). An invariant domain for the compressible Euler equations is the set $\mathcal{A}$ composed of those states with positive density, positive internal energy, and specific entropy $s(\boldsymbol{u})$ larger than $\operatorname{ess\,inf}_{\boldsymbol{x}\in D} s(\boldsymbol{u}_0(\boldsymbol{x}))$. An invariant domain for the compressible Navier–Stokes equations is the set $\mathcal{A}$ composed of those states with positive density and positive internal energy. Another important property of (1) is conservation. Letting $(\boldsymbol{r}_i)_{i\in\{1:m\}}$ be the components of $\boldsymbol{r}$, we assume that there is an index subset $\mathcal{C} \subset \{1{:}m\} := \{1, \dots, m\}$ such that $\boldsymbol{r}_p(\boldsymbol{u}) = 0$ for all $p \in \mathcal{C}$. This means that the relaxation process does not affect the dependent variables indexed in the subset $\mathcal{C}$. Then, again in the absence of perturbations due to the boundary conditions, the following conservation property holds:

$$\int_D \boldsymbol{u}_p(t, \cdot)\,\mathrm{d}x = \int_D \boldsymbol{u}_{0,p}\,\mathrm{d}x, \qquad \forall t \in J,\ \forall p \in \mathcal{C}. \tag{3}$$

The objective of this work is to construct high-order discretizations in space and time that are conservative and leave the set $\mathcal{A}$ invariant. Such methods are called invariant-domain preserving for $\mathcal{A}$, or (IDP) for short. To stay general, our starting point is a system of ordinary differential equations (ODEs) obtained after discretization in space of the conservation equation (1). We mainly focus in this paper on the time discretization. The time discretization methods we are going to present can be combined with various space discretization techniques (e.g., discontinuous and continuous finite elements, finite volumes, finite differences, etc.). We assume that the ODE system takes the following generic form:

$$\mathbb{M}\partial_t \mathbf{U} = \mathbf{F}(\mathbf{U}) + \mathbf{G}(\mathbf{U}),\ \forall t \in J, \qquad \mathbf{U}(0) = \mathbf{U}_0. \tag{4}$$

The mass matrix $\mathbb{M}$ is induced by the space discretization (it is the Gram matrix associated with the global shape functions of the space approximation). The dependent variable $\mathbf{U}(t)$ takes values in $(\mathbb{R}^m)^I$ where $I \geq 1$ is the number of degrees of freedom (dofs) employed in the space discretization. We set $\mathcal{V} := \{1{:}I\} := \{1, \dots, I\}$ and write $\mathbf{U}(t) := (\mathbf{U}_i(t))_{i\in\mathcal{V}}$. For all $i \in \mathcal{V}$, the local state vector $\mathbf{U}_i(t) = (\mathbf{U}_{p,i}(t))_{p\in\{1:m\}}$ is viewed as an approximation of the exact solution $\boldsymbol{u}(t, \cdot)$ at some point in D, say $\boldsymbol{x}_i$. The nonlinear mappings $\mathbf{F} \in C^0(\mathcal{A}^I; (\mathbb{R}^m)^I)$ and $\mathbf{G} \in C^0(\mathcal{A}^I; (\mathbb{R}^m)^I)$ result from the space discretization of the operators $-\boldsymbol{f}$ and $-\boldsymbol{g}$ in (1), respectively, and $\mathbf{U}_0 \in \mathcal{A}^I$ is an appropriate approximation in space of the initial datum $\boldsymbol{u}_0$. We loosely refer to the mappings $\mathbf{F}$ and $\mathbf{G}$ as the hyperbolic flux and the parabolic flux, respectively. Assuming that $\mathbf{U}_i(0) \in \mathcal{A}$ for all $i \in \mathcal{V}$, a natural requirement for the space approximation is that it is invariant-domain preserving, i.e.,

$$\mathbf{U}(t) \in \mathcal{A}^I, \qquad \forall t \in J. \tag{5}$$

A second requirement is that conservation holds:

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}(t) = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}(0), \qquad \forall t \in J,\ \forall p \in \mathcal{C}. \tag{6}$$

Using an implicit scheme to discretize in time the ODE system (4) is often too expensive owing to the nonlinearities involved in the fluxes $\mathbf{F}$ and $\mathbf{G}$, whereas using an explicit scheme results in a severe restriction on the time step owing to stiffness induced by the parabolic flux $\mathbf{G}$. A well known remedy to this conundrum is to resort to implicit-explicit (IMEX) schemes, where the numerical flux $\mathbf{F}$ is treated explicitly, and the numerical flux $\mathbf{G}$ is treated implicitly. The origin of IMEX schemes can be traced back to Crouzeix [9], Varah [28]. We also refer the reader to Ascher et al. [1, 2], Burman and Ern [5], Kennedy and Carpenter [17], Pareschi and Russo [24, 25], Zhong [32] for other developments. Despite these advances, a crucial question that still remains open is how to reconcile the use of an IMEX time stepping scheme with the above invariant-domain property, while at the same time ensuring conservation. Building on [10], we propose an answer to this question in this paper. More precisely, we introduce a technique that makes every IMEX Runge–Kutta (RK) time stepping method invariant-domain preserving (IDP) and conservative. The resulting schemes are called "IDP-IMEX" schemes.

This work is organized as follows. In Section 2, we outline the discrete setting in space and time, we identify the key assumptions underlying this work, and we exemplify these notions for the Euler IDP-IMEX scheme. In Section 3, we extend these ideas and build higher-order IDP-IMEX schemes. We introduce a generic IDP-IMEX algorithm composed of the steps (51) to (61) whose properties are stated in Theorem 3.3. In Section 4, we review some examples of higher-order IMEX schemes and we derive some novel examples with optimal efficiency. Finally, in Section 5, we present numerical illustrations on stiff ODEs and a nonlinear scalar conservation equation.

2 Preliminaries

The goal of this section is threefold: (i) introduce the discrete setting in space and time; (ii) identify the key ideas and assumptions underlying this work; (iii) exemplify these notions for the Euler IDP-IMEX scheme. All this material is used in Section 3 where we introduce the novel higher-order IDP-IMEX schemes.

2.1 Time discretization and quasi-linearization

Let $t^n \in [0,T]$ be the current time with $n \in \{0{:}N\}$, $t^0 := 0$, and $t^N := T$. Let τ^n be the current time step and let $t^{n+1} := t^n + \tau^n$. To simplify the notation, we henceforth write τ instead of τ^n. Let $\mathbf{U}^n$ be the approximation of the solution to (4) at the discrete time t^n. The key invariant-domain property we want to achieve is the following:

$$(\mathbf{U}^n \in \mathcal{A}^I) \Longrightarrow (\mathbf{U}^{n+1} \in \mathcal{A}^I), \qquad \forall n \geq 0. \tag{7}$$

Moreover, we want to achieve this goal while maintaining conservation. The notion of conservation will be made more precise below, but for the time being, conservation is expressed at the global level by requiring that

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n}, \qquad \forall p \in \mathcal{C}. \tag{8}$$

To avoid solving a nonlinear problem at each time-step when the parabolic fluxes are made implicit, we introduce a quasi-linearization process (the way this is done is made more precise in

§2.2 and §2.3). We consider a quasi-linearized parabolic flux $\mathbf{G}^{\rm lin} \in C^0(\mathcal{A}^I \times (\mathbb{R}^m)^I; (\mathbb{R}^m)^I)$ that is consistent with $\mathbf{G}$, i.e.,

$$\mathbf{G}^{\rm lin}(\mathbf{W};\mathbf{W}) = \mathbf{G}(\mathbf{W}), \quad \forall \mathbf{W} \in \mathcal{A}^I. \tag{9}$$

We assume that this flux is such that for all $\mathbf{W} \in \mathcal{A}^I$, the problem consisting of seeking $\mathbf{U} \in (\mathbb{R}^m)^I$ so that

$$\mathbb{M}\mathbf{U} - \tau \mathbf{G}^{\rm lin}(\mathbf{W};\mathbf{U}) = \mathbb{M}\mathbf{W} \tag{10}$$

is well-posed and easy to solve. For instance, this problem could only involve linear solves. Notice that this does not mean that the mapping $\mathbf{U} \mapsto \mathbf{G}(\mathbf{W};\mathbf{U})$ is linear; see §5.1 for an example. Owing to the above quasi-linearization process, we reformulate (4) over the time interval $J^n := [t^n, t^{n+1}]$ as follows: Find $\mathbf{U} \in C^1(J^n; (\mathbb{R}^m)^I)$ so that $\mathbf{U}(t^n) = \mathbf{U}^n$ and for all $t \in J^n$,

$$\mathbb{M}\partial_t \mathbf{U} = \underbrace{\mathbf{F}(\mathbf{U}) + \mathbf{G}(\mathbf{U}) - \mathbf{G}^{\rm lin}(\mathbf{U}^n;\mathbf{U})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\rm lin}(\mathbf{U}^n;\mathbf{U})}_{\text{implicit}}. \tag{11}$$

2.2 Space discretization and conservation structure

Let us now give details on the space discretization. We consider two space discretizations. The first one is low-order accurate and referred to with the superscript $^{\rm L}$. The second one is high-order accurate and referred to with the superscript $^{\rm H}$. The low-order scheme is based on a low-order invertible mass matrix $\mathbb{M}^{\rm L} \in \mathbb{R}^{I\times I}$ and low-order fluxes $\mathbf{F}^{\rm L}, \mathbf{G}^{\rm L} : \mathcal{A}^I \to (\mathbb{R}^m)^I$. The high-order scheme is based on a high-order invertible mass matrix $\mathbb{M}^{\rm H} \in \mathbb{R}^{I\times I}$ and high-order fluxes $\mathbf{F}^{\rm H}, \mathbf{G}^{\rm H} : \mathcal{A}^I \to (\mathbb{R}^m)^I$.

We assume that $\mathbb{M}^{\rm H}$ is symmetric positive-definite with entries $(m_{ij})_{i,j\in\mathcal{V}}$ and $\mathbb{M}^{\rm L}$ is diagonal with entries $(\delta_{ij}m_i)_{i,j\in\mathcal{V}}$. For all $i \in \mathcal{V}$, we introduce the subset $\mathcal{I}(i) \subsetneq \mathcal{V}$ such that $m_{ij} \neq 0$ for all $j \in \mathcal{I}(i)$. We call $\mathcal{I}(i)$ stencil at i. The notion of stencil is symmetric, i.e., $j \in \mathcal{I}(i)$ if and only if $i \in \mathcal{I}(j)$ because $m_{ij} = m_{ji}$. Finally, we assume that

$$m_i = \sum_{j\in\mathcal{V}} m_{ij} = \sum_{j\in\mathcal{V}} m_{ji}, \qquad \forall i \in \mathcal{V}, \tag{12}$$

In the finite element terminology, this means that the low-order mass matrix $\mathbb{M}^{\rm L}$ is the lumped version of the high-order mass matrix $\mathbb{M}^{\rm H}$. For every matrix $\mathbb{M} \in \mathbb{R}^{I\times I}$ and every vector $\mathbf{V} \in (\mathbb{R}^m)^I$ with components $\mathbf{V}_{p,i}$, with $p \in \{1{:}m\}$ and $i \in \mathcal{V}$, the components of the vector $\mathbb{M}\mathbf{V} \in (\mathbb{R}^m)^I$ are defined to be $(\mathbb{M}\mathbf{V})_{p,i} := \sum_{j\in\mathcal{V}} m_{ij}\mathbf{V}_{p,j}$ for all $p \in \{1{:}m\}$ and $i \in \mathcal{V}$.

The components of the low-order and high-order hyperbolic fluxes are denoted $\mathbf{F}_i^{\rm L}(\mathbf{V}) \in \mathbb{R}^m$ and $\mathbf{F}_i^{\rm H}(\mathbf{V}) \in \mathbb{R}^m$ for all $i \in \mathcal{V}$ and all $\mathbf{V} \in \mathcal{A}^I$. To account for the fact that the hyperbolic fluxes are associated with a conservation principle, we assume that these fluxes admit the following stencil-based decomposition:

$$\mathbf{F}_i^{\rm L}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\rm L}(\mathbf{V}), \qquad \mathbf{F}_i^{\rm H}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\rm H}(\mathbf{V}), \qquad \forall \mathbf{V} \in \mathcal{A}^I, \tag{13}$$

where $\mathbf{F}_{ij}^{\rm L}, \mathbf{F}_{ij}^{\rm H} \in C^0(\mathcal{A}^I; \mathbb{R}^m)$, and we assume the following skew-symmetry property:

$$\mathbf{F}_{ij}^{\rm L}(\mathbf{V}) = -\mathbf{F}_{ji}^{\rm L}(\mathbf{V}), \qquad \mathbf{F}_{ij}^{\rm H}(\mathbf{V}) = -\mathbf{F}_{ji}^{\rm H}(\mathbf{V}), \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{14}$$

The same structure is assumed for the parabolic fluxes, namely (for brevity, we only write the statements for the high-order fluxes)

$$\mathbf{G}_i^{\rm H}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}_{ij}^{\rm H}(\mathbf{V}) + \mathbf{R}_i^{\rm H}(\mathbf{V}), \qquad \mathbf{D}_{ij}^{\rm H}(\mathbf{V}) = -\mathbf{D}_{ji}^{\rm H}(\mathbf{V}),\ \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{15}$$

Consistently with our assumption that $\boldsymbol{r}_p(\boldsymbol{u}) = 0$ for all $p \in \mathcal{C}$, we assume that $\mathbf{R}^{\mathrm{H}}_{p,i}(\mathbf{V}) = 0$ for all $p \in \mathcal{C}$ and all $i \in \mathcal{V}$.

The quasi-linearization process mentioned in (11) is performed for both the low-order and high-order parabolic fluxes. This leads to quasi-linearized parabolic fluxes $\mathbf{G}^{\mathrm{L,lin}}, \mathbf{G}^{\mathrm{H,lin}} \in C^0(\mathcal{A}^I \times (\mathbb{R}^m)^I; (\mathbb{R}^m)^I)$ which we assume satisfy the following decompositions and properties:

$$\mathbf{G}^{\mathrm{L,lin}}_i(\mathbf{W};\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}^{\mathrm{L,lin}}_{ij}(\mathbf{W};\mathbf{V}) + \mathbf{R}^{\mathrm{L,lin}}_i(\mathbf{W};\mathbf{V}), \tag{16}$$

$$\mathbf{G}^{\mathrm{H,lin}}_i(\mathbf{W};\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}^{\mathrm{H,lin}}_{ij}(\mathbf{W};\mathbf{V}) + \mathbf{R}^{\mathrm{H,lin}}_i(\mathbf{W};\mathbf{V}), \tag{17}$$

$$\mathbf{D}^{\mathrm{L,lin}}_{ij}(\mathbf{W};\mathbf{V}) = -\mathbf{D}^{\mathrm{L,lin}}_{ji}(\mathbf{W};\mathbf{V}), \qquad \mathbf{D}^{\mathrm{H,lin}}_{ij}(\mathbf{W};\mathbf{V}) = -\mathbf{D}^{\mathrm{H,lin}}_{ji}(\mathbf{W};\mathbf{V}), \tag{18}$$

$$\mathbf{R}^{\mathrm{L,lin}}_{p,i} = \mathbf{R}^{\mathrm{H,lin}}_{p,i} = 0, \qquad \forall p \in \mathcal{C},\ \forall i \in \mathcal{V}. \tag{19}$$

In conclusion, we are going to consider two versions of the ODE system (11) over the time interval $J^n = [t^n, t^{n+1}]$. One corresponds to the low-order space discretization:

$$\mathbb{M}^{\mathrm{L}}\partial_t \mathbf{U}^{\mathrm{L}} = \underbrace{\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{\mathrm{L}})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\mathrm{L,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{L}})}_{\text{implicit}}. \tag{20}$$

The other one corresponds to the high-order space discretization:

$$\mathbb{M}^{\mathrm{H}}\partial_t \mathbf{U}^{\mathrm{H}} = \underbrace{\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{\mathrm{H}}) + \mathbf{G}^{\mathrm{H}}(\mathbf{U}^{\mathrm{H}}) - \mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H}})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H}})}_{\text{implicit}}. \tag{21}$$

Consistently with our assumption on (10), we assume that for all $\mathbf{V} \in (\mathbb{R}^m)^I$ and all $\mathbf{W} \in \mathcal{A}^I$, the problems consisting of seeking $\mathbf{U}^{\mathrm{L}}, \mathbf{U}^{\mathrm{H}} \in (\mathbb{R}^m)^I$ so that

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L}} - \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W};\mathbf{U}^{\mathrm{L}}) = \mathbb{M}^{\mathrm{L}}\mathbf{V}, \qquad \mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H}} - \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{W};\mathbf{U}^{\mathrm{H}}) = \mathbb{M}^{\mathrm{H}}\mathbf{V}, \tag{22}$$

are well-posed and easy to solve.

2.3 Structural IDP assumptions

To gently introduce our ideas, let us consider the well-known IMEX method consisting of combining the forward and the backward Euler time steppings. We call this method Euler IMEX. Let us apply the method to the low-order ODE system (20). Let $\mathbf{U}^n \in \mathcal{A}^I$. The first step is explicit and consists of computing the hyperbolic prediction

$$\mathbf{W}^{\mathrm{L},n} := \big(\mathbb{I} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}\big)(\mathbf{U}^n). \tag{23}$$

The second step is implicit and consists of computing the final state $\mathbf{U}^{\mathrm{L},n+1}$ by solving the quasi-linear problem

$$\big(\mathbb{I} - \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{\mathrm{L},n};\cdot)\big)(\mathbf{U}^{\mathrm{L},n+1}) = \mathbf{W}^{\mathrm{L},n}. \tag{24}$$

Altogether, we have

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} = \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n) + \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{\mathrm{L},n};\mathbf{U}^{\mathrm{L},n+1}). \tag{25}$$

This leads us to formulate the following two key structural assumptions on the low-order fluxes:

Assumption 2.1 (Low-order fluxes)**.** *There exists $\tau^* > 0$ s.t. for all $\tau \in (0, \tau^*]$,*

(i) *the low-order hyperbolic flux satisfies the following property:*

$$\{\, \mathbf{V} \in \mathcal{A}^I \,\} \Longrightarrow \{\, \big(\mathbb{I} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}\big)(\mathbf{V}) \in \mathcal{A}^I \,\}. \tag{26}$$

(ii) *the low-order quasi-linearized parabolic flux satisfies the following property:*

$$\{\, \mathbf{V} \in \mathcal{A}^I \,\} \Longrightarrow \{\, \big(\mathbb{I} - \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{G}^{\mathrm{L,lin}}(\mathbf{V};\cdot)\big)^{-1}(\mathbf{V}) \in \mathcal{A}^I \,\}. \tag{27}$$

The following result prefigures what we are aiming at. We omit the proof since it is somewhat standard.

Lemma 2.2 (Low-order Euler IDP-IMEX scheme)**.** *Assume that* $\mathbf{U}^n \in \mathcal{A}^I$ *and* $\tau \in (0, \tau^*]$*. Then, the low-order Euler IMEX scheme* (25) *is well-defined. Moreover, it is IDP and conservative, i.e.,* $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$ *and* $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n}$ *for all* $p \in \mathcal{C}$*.*

Remark 2.3 (Time step restriction)**.** *In many situations, the time step restriction* $\tau \in (0, \tau^*]$ *is only required for the invariant-domain property of the hyperbolic step* (26)*. The invariant-domain property of the parabolic step* (27) *can often be shown to hold for every time step* $\tau > 0$*.*

Of course, the above result is of little interest since what we actually want is to use a high-order approximation in space. The Euler IMEX scheme applied to the high-order ODE system (21) consists of seeking $\mathbf{U}^{\mathrm{H},n+1} \in (\mathbb{R}^m)^I$ so that

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n} + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n) + \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}). \tag{28}$$

Similarly to (25), the method (28) is composed of two steps. The first one consists of computing the forward Euler prediction $\mathbf{W}^{\mathrm{H},n} := \big(\mathbb{I}+\tau(\mathbb{M}^{\mathrm{H}})^{-1}\mathbf{F}^{\mathrm{H}}\big)(\mathbf{U}^n)$. The second one consists of computing the parabolic update $\mathbf{U}^{\mathrm{H},n+1}$ by solving the quasi-linear problem $\big(\mathbb{I}-\tau(\mathbb{M}^{\mathrm{H}})^{-1}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\cdot)\big)(\mathbf{U}^{\mathrm{H},n+1}) = \mathbf{W}^{\mathrm{H},n}$. Unfortunately, even though we are using forward and backward Euler time stepping, there is no guarantee that $\mathbf{U}^{\mathrm{H},n+1}$ belongs to $\mathcal{A}^I$, i.e., the high-order counterpart of Lemma 2.2 does not hold true in general.

This problem is solved in the literature by using nonlinear limiting operators; see Boris and Book [3], Harten [16], Osher and Chakravarthy [23], Zalesak [29]. Limiting is realized in the discontinuous Galerkin and finite volume settings by squeezing the high-order approximation towards the piecewise constant approximation over each mesh cell (see Sanders [26, Thm. 2.1], Coquel and LeFloch [7, Thm .4.3], Liu and Osher [21, Thm. 1], and Zhang and Shu [30, Thm. 2.5]). A well-known limiting method for scalar conservation equations is the so-called flux-corrected transport (FCT) technique of Boris and Book [3] and Zalesak [29]. The reader is also referred to Kuzmin and Turek [19] and Kuzmin et al. [20] for other extensions on this method in the context of finite elements. When the bounds to be enforced are non-affine (as is often the case for hyperbolic systems), one has to use nonlinear methods like in, e.g., [26, Lem. 3.3], [7, Thm .4.3], [21, Thm. 2], or Zhang and Shu [31, Lem. 2.4], or other nonlinear variants like convex limiting (see, e.g., Guermond et al. [12, 13]). The key idea common to all the above techniques is the decomposition of the flux over the stencils in skew-symmetric components as in (13), (14), and (15).

In the present work, we are going to consider two limiters: one to compute the hyperbolic prediction and another to compute the parabolic update. We now introduce the corresponding notation. Let $\mathfrak{L}$ be the collection of the sparse symmetric matrices with coefficients in $[0, 1]$ and with the sparsity pattern induced by the stencils $(\mathcal{I}(i))_{i\in\mathcal{V}}$. We also let $\mathfrak{M}$ be the collection of the skew-symmetric matrices in $(\mathbb{R}^m)^{I\times I}$ with the block-sparsity pattern induced by the stencils $(\mathcal{I}(i))_{i\in\mathcal{V}}$. Finally, we define $\mathcal{B} := \{\boldsymbol{z} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \in \mathbb{R}^m \mid \boldsymbol{z}_p = 0, \forall p \in \mathcal{C}\}$.

Definition 2.4 (Conservative hyperbolic limiter)**.** *Let* $\big(\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\mathbf{A}_{ij}\big)_{i\in\mathcal{V}}$ *be an hyperbolic prediction, with* $\mathbf{V} := (\mathbf{V}_i)_{i\in\mathcal{V}} \in \mathcal{A}^I$ *and* $\mathbf{A} := (\mathbf{A}_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)} \in \mathfrak{M}$*. We call conservative hyperbolic limiter any operator* $\ell^{\rm hyp} : \mathcal{A}^I\times\mathfrak{M} \ni (\mathbf{V},\mathbf{A}) \mapsto (\ell_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)} \in \mathfrak{L}$ *s.t. the following holds:*

$$\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij} \in \mathcal{A}, \quad \forall i \in \mathcal{V}. \tag{29}$$

For brevity, the state $(\mathbf{V}_i + m_i^{-1}\tau\sum_{j\in\mathcal{I}(i)}\ell_{ij}\mathbf{A}_{ij})_{i\in\mathcal{V}} \in \mathcal{A}^I$ *is denoted* $\boldsymbol{\ell}^{\rm hyp}(\mathbf{V},\mathbf{A})$*.*

Definition 2.5 (Conservative parabolic limiter)**.** *Let* $\big(\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\mathbf{A}_{ij} + \frac{\tau}{m_i}\mathbf{B}_i\big)_{i\in\mathcal{V}}$ *be a parabolic update with* $\mathbf{V} := (\mathbf{V}_i)_{i\in\mathcal{V}} \in \mathcal{A}^I$*,* $\mathbf{A} := (\mathbf{A}_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)} \in \mathfrak{M}$*, and* $\mathbf{B} := (\mathbf{B}_i)_{i\in\mathcal{V}} \in \mathcal{B}^I$*. We call conservative parabolic limiter any operator* $\ell^{\rm par} : \mathcal{A}^I\times\mathfrak{M}\times\mathcal{B}^I \ni (\mathbf{V},\mathbf{A},\mathbf{B}) \mapsto \big((\ell^a_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)}, (\ell^b_i)_{i\in\mathcal{V}}\big) \in \mathfrak{L}\times[0,1]^I$ *s.t. the following holds:*

$$\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \ell^a_{ij}\mathbf{A}_{ij} + \frac{\tau}{m_i}\ell^b_i\mathbf{B}_i \in \mathcal{A}, \quad \forall i \in \mathcal{V}. \tag{30}$$

For brevity, the state $\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\ell^a_{ij}\mathbf{A}_{ij} + \frac{\tau}{m_i}\ell^b_i\mathbf{B}_i$ *is denoted* $\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})$*.*

The existence of limiters is guaranteed since the trivial limiters $\boldsymbol{\ell}^{\rm hyp}(\mathbf{V},\mathbf{A}) = \mathbf{V}$ (i.e., $\ell_{ij} = 0$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$) and $\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B}) = \mathbf{V}$ (i.e., $\ell^a_{ij} = \ell^b_i = 0$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$) are always admissible because $\mathbf{V} \in \mathcal{A}^I$. Of course, the trivial limiters are inefficient. The goal of limiters is to construct the limiting coefficients ℓ_{ij}, ℓ^a_{ij} and ℓ^b_i as close to 1 as possible. Regardless of the values taken by the limiters, an important property of the limiters is conservativity.

Lemma 2.6 (Conservation)**.** *For all* $(\mathbf{V},\mathbf{A},\mathbf{B}) \in \mathcal{A}^I\times\mathfrak{M}\times\mathcal{B}^I$*, all* $p \in \mathcal{C}$*, we have*

$$\sum_{i\in\mathcal{V}} m_i\boldsymbol{\ell}^{\rm hyp}(\mathbf{V},\mathbf{A})_{p,i} = \sum_{i\in\mathcal{V}} m_i\mathbf{V}_{p,i}, \qquad \sum_{i\in\mathcal{V}} m_i\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})_{p,i} = \sum_{i\in\mathcal{V}} m_i\mathbf{V}_{p,i}. \tag{31}$$

Proof. For the parabolic limiter, we have $\sum_{i\in\mathcal{V}} m_i\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})_{p,i} = \sum_{i\in\mathcal{V}} m_i\mathbf{V}_i + \tau\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)}\ell^a_{ij}\mathbf{A}_{ij}$ for all $p \in \mathcal{C}$ because $\mathbf{B} \in \mathcal{B}^I$. But the symmetry property $\ell^a_{ij} = \ell^a_{ji}$ and the skew-symmetry property $\mathbf{A}_{ij} = -\mathbf{A}_{ji}$ imply that $\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)}\ell^a_{ij}\mathbf{A}_{ij} = 0$, whence the assertion. The proof for the hyperbolic limiter is similar. □

2.4 Euler IDP-IMEX scheme

All the ingredients are now in place to make the Euler IMEX scheme invariant-domain preserving with high-order space discretization. Given $\mathbf{U}^n \in \mathcal{A}^I$, the scheme is decomposed into the following four steps:

$$\mathbf{U}^n \underbrace{\xrightarrow{(1)} (\mathbf{W}^{\mathrm{L},n+1},\mathbf{W}^{\mathrm{H},n+1}) \xrightarrow{(2)}}_{\text{hyperbolic step}} \mathbf{W}^{n+1} \underbrace{\xrightarrow{(3)} (\mathbf{U}^{\mathrm{L},n+1},\mathbf{U}^{\mathrm{H},n+1}) \xrightarrow{(4)}}_{\text{parabolic step}} \mathbf{U}^{n+1}. \tag{32}$$

Let us now give the details of the four steps. We essentially follow Zalesak's limiting strategy [29, Eq. (4)] for both the hyperbolic and the parabolic steps.

Hyperbolic steps (1) and (2)

Step (1) consists of computing the low-order and high-order hyperbolic updates

$$\mathbb{M}^{\mathrm{L}}\mathbf{W}^{\mathrm{L},n+1} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n), \tag{33}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{\mathrm{H},n+1} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n). \tag{34}$$

In Step (2), we apply the hyperbolic limiting operator. Subtracting (33) from (34) and using (13) and (14), elementary manipulations show that for all $i \in \mathcal{V}$,

$$\mathbf{W}_i^{\mathrm{H},n+1} = \mathbf{W}_i^{\mathrm{L},n+1} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\mathbf{A}_{ij}^n, \tag{35}$$

with

$$\mathbf{A}_{ij}^n := \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n) - \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n) + \frac{m_i\delta_{ij} - m_{ij}}{\tau}(\mathbf{W}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n - \mathbf{W}_i^{\mathrm{H},n+1} + \mathbf{U}_i^n). \tag{36}$$

Notice that $\mathbf{A}^n$ is indeed skew-symmetric. Then using Definition 2.4, the conservative IDP hyperbolic high-order update is obtained by setting

$$\mathbf{W}^{n+1} := \boldsymbol{\ell}^{\mathrm{hyp}}(\mathbf{W}^{\mathrm{L},n+1}, \mathbf{A}^n). \tag{37}$$

Parabolic steps (3) and (4)

Step (3) consists of computing the low-order and high-order parabolic updates by solving

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} - \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}) := \mathbb{M}^{\mathrm{L}}\mathbf{W}^{n+1}, \tag{38}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} - \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},n+1}) := \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n+1}. \tag{39}$$

The quasi-linearization in (38) is based on $\mathbf{W}^{n+1}$ to invoke Assumption (27). The quasi-linearization in (39) is based on $\mathbf{U}^n$ to be consistent with the higher-order case to be explained in the next section (see also Remark 3.1). In Step (4), we apply the parabolic limiting operator. Subtracting (38) from (39) and using (16), (17), and (18), elementary manipulations show that for all $i \in \mathcal{V}$,

$$\mathbf{U}_i^{\mathrm{H},n+1} = \mathbf{U}_i^{\mathrm{L},n+1} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\mathbf{A}_{ij}^n + \frac{\tau}{m_i}\mathbf{B}_i^n, \tag{40}$$

with

$$\begin{aligned}\mathbf{A}_{ij}^n := {}& \mathbf{D}_{ij}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}) - \mathbf{D}_{ij}^{\mathrm{L,lin}}(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}) \\ & + \tau^{-1}(m_i\delta_{ij} - m_{ij})(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{W}_j^{n+1} - \mathbf{U}_i^{\mathrm{H},n+1} + \mathbf{W}_i^{n+1}),\end{aligned} \tag{41}$$

$$\mathbf{B}_i^n := \mathbf{R}_i^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}) - \mathbf{R}_i^{\mathrm{L,lin}}(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}). \tag{42}$$

Notice that $\mathbf{A}^n$ is indeed skew-symmetric and $\mathbf{B}^n \in \mathcal{B}^I$. Then using Definition 2.5, the conservative IDP parabolic high-order update is obtained by setting

$$\mathbf{U}^{n+1} := \boldsymbol{\ell}^{\mathrm{par}}(\mathbf{U}^{\mathrm{L},n+1}, \mathbf{A}^n, \mathbf{B}^n). \tag{43}$$

Conclusion

The main result for the above construction is the following.

Lemma 2.7 (High-order Euler IDP-IMEX scheme)**.** *Let Assumption* 2.1 *hold and assume* $\tau \in (0, \tau^*]$*. Assume that the limiters match Definitions* 2.4-2.5*. Let* $\mathbf{U}^n \in \mathcal{A}^I$*. Then, the above Euler IMEX schemes* (33)–(43) *satisfies the following properties:*

(i) *It is well defined;*

(ii) *It is IDP, i.e., it satisfies* (7)*;*

(iii) *It is conservative, i.e., it satisfies* (8)*.*

Proof. (1) Let us first prove the assertions (i) and (ii). Since $\mathbf{U}^n \in \mathcal{A}^I$, invoking Assumption 2.1(i) gives $\mathbf{W}^{\mathrm{L},n+1} \in \mathcal{A}^I$. The definition of the hyperbolic limiter then implies that $\mathbf{W}^{n+1} \in \mathcal{A}^I$. Invoking Assumption 2.1(ii) then shows that $\mathbf{U}^{\mathrm{L},n+1}$ is well-defined and $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$. Finally, the definition of the parabolic limiter implies that $\mathbf{U}^{n+1} \in \mathcal{A}^I$.
(2) Conservativity follows from the following identities for all $p \in \mathcal{C}$:

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{W}_{p,i}^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{W}_{p,i}^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n},$$

where the first and third equalities follow from Lemma 2.6, whereas the second and fourth equalities follow from the skew-symmetry assumption on the low-order fluxes. □

3 High-order IDP-IMEX schemes

In this section, we extend the construction described in §2.4 to any IMEX scheme combining an explicit Runge–Kutta (ERK) scheme with a diagonally implicit RK scheme whose first stage is fully explicit (EDIRK). We assume that both schemes consist of s stages, $s \geq 2$. The main original ideas of this paper are in this section.

3.1 Butcher tableaux

The ERK and EDIRK schemes are described by their respective Butcher tableau which we assume to have the following form:

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a^{\mathsf{e}}_{2,1} & 0 & & & \\ c_3 & a^{\mathsf{e}}_{3,1} & a^{\mathsf{e}}_{3,2} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ c_s & a^{\mathsf{e}}_{s,1} & a^{\mathsf{e}}_{s,2} & \cdots & a^{\mathsf{e}}_{s,s-1} & 0 \\ \hline & b^{\mathsf{e}}_1 & b^{\mathsf{e}}_2 & \cdots & b^{\mathsf{e}}_{s-1} & b^{\mathsf{e}}_s \end{array} \qquad \begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a^{\mathsf{i}}_{2,1} & a^{\mathsf{i}}_{2,2} & & & \\ c_3 & a^{\mathsf{i}}_{3,1} & a^{\mathsf{i}}_{3,2} & a^{\mathsf{i}}_{3,3} & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ c_s & a^{\mathsf{i}}_{s,1} & a^{\mathsf{i}}_{s,2} & \cdots & a^{\mathsf{i}}_{s,s-1} & a^{\mathsf{i}}_{s,s} \\ \hline & b^{\mathsf{i}}_1 & b^{\mathsf{i}}_2 & \cdots & b^{\mathsf{i}}_{s-1} & b^{\mathsf{i}}_s \end{array} \tag{44}$$

Notice that $a^{\mathsf{e}}_{l,k} := 0$ for all $k \geq l$, $a^{\mathsf{i}}_{l,k} := 0$ for all $k > l$. Here, the superscript $^{\mathsf{e}}$ refers to the ERK scheme and the superscript $^{\mathsf{i}}$ refers to the EDIRK scheme. Recall that the coefficients c_j define the intermediate time steps $t^{n,j} := t^n + c_j\tau$. Notice that both schemes share the same set of coefficients $(c_j)_{j\in\{1:s\}}$. For convenience, we set $c_{s+1} := 1$. We assume that $c_1 = 0$ and $c_j \geq 0$ for all $j \in \{2{:}s\}$. To simplify some expressions, we set $a^{\mathsf{e}}_{s+1,j} := b^{\mathsf{e}}_j$ and $a^{\mathsf{i}}_{s+1,j} := b^{\mathsf{i}}_j$ for all $j \in \{1{:}s\}$. Whenever $a^{\mathsf{i}}_{i,1} = 0$ for all $i \in \{1{:}s+1\}$, the EDIRK scheme is called zero-padded. If all the diagonal entries

$a^{\mathsf{i}}_{i,i}$ are equal (except the first one which is zero), we speak of singly diagonal EDIRK scheme (ESDIRK).

We are going to assume that

$$\sum_{l\in\{1:j\}} a^{\mathsf{e}}_{j,l} = \sum_{l\in\{1:j\}} a^{\mathsf{i}}_{j,l} = c_j, \qquad \forall j \in \{1{:}s\}. \tag{45}$$

This is one of Butcher's simplifying assumptions for each RK scheme. This assumption implies that $a^{\mathsf{e}}_{1,1} = a^{\mathsf{i}}_{1,1} = 0$. Moreover, since consistency requires that $\sum_{j\in\{1:s\}} b^{\mathsf{e}}_j = \sum_{j\in\{1:s\}} b^{\mathsf{i}}_j = 1$, the identity (45) also holds true for $j = s+1$. Finally, we use the convention that $a^{\mathsf{i}}_{s+1,s+1} := 0$.

3.2 High-order IMEX scheme in incremental form

Following the ideas developed in [10] for ERK schemes, we write the IMEX scheme in incremental form. To this purpose, for all $l \in \{2{:}s+1\}$, we define the stage index $l'(l)$ to be the largest index in $\{1{:}l-1\}$ so that $c_l - c_{l'(l)}$ is the minimal value of $c_l - c_k$ for all $k \in \{1{:}l-1\}$ so that $c_l - c_k \geq 0$:

$$l'(l) := \min\{k \in \{1{:}l-1\} \mid c_l - c_k \geq 0\}, \qquad \forall l \in \{2{:}s+1\}. \tag{46}$$

Owing to the assumption $c_l \geq 0 = c_1$ for all $l \in \{2{:}s+1\}$, we infer that $1 \in \{k \in \{1{:}l-1\} \mid c_l - c_k \geq 0\}$, which means that the set $\{k \in \{1{:}l-1\} \mid c_l - c_k \geq 0\}$ is nonempty and the above definition makes sense. The definition of $l'(l)$ remains meaningful for so-called confluent RK methods for which several c_l's take the same value. If the sequence $(c_l)_{l\in\{1:s\}}$ is nondecreasing, then $l'(l) = l-1$ for all $l \in \{2{:}s+1\}$. The reason for looking for the smallest difference $c_l - c_{l'(l)}$ is to minimize the CFL restriction on the time step (see Assumption 2.1(i)). For further reference, we define

$$\Delta c^{\max} := \max_{2\leq l\leq s+1} \big(c_l - c_{l'(l)}\big). \tag{47}$$

Notice that $\Delta c^{\max} \geq \frac{1}{s}$ and $\Delta c^{\max} = \frac{1}{s}$ whenever all the stages of the ERK method are equidistributed, i.e., $c_l = \frac{l-1}{s}$, $l \in \{1{:}s+1\}$. In the rest of this paper, we simply write l' instead of $l'(l)$ to simplify the notation.

We can now approximate in time the high-order ODE system (21) by using the IMEX method defined by the two Butcher tableaux in (44). We first set $\mathbf{U}^{n,1} := \mathbf{U}^n$. Then, for all $l \in \{2{:}s+1\}$, the l-th stage of the IMEX scheme consists of computing the following high-order update:

$$\begin{aligned}\mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} &+ \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\big(\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}) + \mathbf{G}^{\mathrm{H}}(\mathbf{U}^{n,k}) - \mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k})\big)\\ &+ \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,l}) + \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{i}}_{l,k} - a^{\mathsf{i}}_{l',k})\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k}).\end{aligned}$$

This incremental form of the IMEX scheme is obtained by subtracting the equation defining the IMEX update at stage l' from the equation defining the update at stage l. We decompose the above problem into one hyperbolic prediction followed by one parabolic update as follows:

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}), \tag{48}$$

$$\begin{aligned}\mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l} - \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,l}) := \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} &+ \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{i}}_{l,k} - a^{\mathsf{i}}_{l',k})\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k})\\ &+ \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\big(\mathbf{G}^{\mathrm{H}}(\mathbf{U}^{n,k}) - \mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k})\big).\end{aligned} \tag{49}$$

Notice that, following (11), the quasi-linearization process uses the initial state $\mathbf{U}^n$ at all stages. We now explain how to make the method (48)-(49) IDP.

3.3 IDP-IMEX scheme

We proceed as in §2.4 to make the scheme (48)-(49) invariant-domain preserving. Given $\mathbf{U}^n \in \mathcal{A}^I$, we set $\mathbf{U}^{n,1} := \mathbf{U}^n$ and we decompose each stage $l \in \{2{:}s+1\}$ into the following four steps:

$$\mathbf{U}^{n,l'} \underbrace{\xrightarrow{(1)} (\mathbf{W}^{\mathrm{L},l}, \mathbf{W}^{\mathrm{H},l}) \xrightarrow{(2)}}_{\text{hyperbolic step (48)}} \mathbf{W}^{n,l} \underbrace{\xrightarrow{(3)} (\mathbf{U}^{\mathrm{L},l}, \mathbf{U}^{\mathrm{H},l}) \xrightarrow{(4)}}_{\text{parabolic step (49)}} \mathbf{U}^{n,l}. \tag{50}$$

Hyperbolic steps (1) and (2)

The IDP realization of the hyperbolic update (48) is done as in [10]. One first computes the low-order and high-order hyperbolic updates defined by

$$\mathbb{M}^{\mathrm{L}}\mathbf{W}^{\mathrm{L},l} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \tag{51}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{52}$$

By proceeding as in (35)-(36), we have

$$\mathbf{W}_i^{\mathrm{H},l} = \mathbf{W}_i^{\mathrm{L},l} + \frac{\tau}{m_i} \sum_{j\in\mathcal{I}(i)} \mathbf{A}_{ij}^{n,l}, \qquad \forall i \in \mathcal{V}, \tag{53}$$

$$\begin{aligned} \text{with} \quad \mathbf{A}_{ij}^{n,l} := & \sum_{k\in\{1:l-1\}} (a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\mathbf{F}^{\mathrm{H}}_{ij}(\mathbf{U}^{n,k}) - (c_l - c_{l'})\mathbf{F}^{\mathrm{L}}_{ij}(\mathbf{U}^{n,l'}) \\ & - \tau^{-1}(m_{ij} - m_i\delta_{ij})(\mathbf{W}_j^{\mathrm{H},l} - \mathbf{U}_j^{n,l'} - \mathbf{W}_i^{\mathrm{H},l} + \mathbf{U}_i^{n,l'}). \end{aligned} \tag{54}$$

Notice that $\mathbf{A}^{n,l}$ is skew-symmetric in compliance with Definition 2.4. Using the hyperbolic limiter, we then set

$$\mathbf{W}^{n,l} := \boldsymbol{\ell}^{\mathrm{hyp}}(\mathbf{W}^{\mathrm{L},l}, \mathbf{A}^{n,l}). \tag{55}$$

Parabolic steps (3) and (4)

We now compute the low-order and high-order parabolic updates defined by solving the following two problems:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} - \tau(c_l - c_{l'})\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{n,l}; \mathbf{U}^{\mathrm{L},l}) := \mathbb{M}^{\mathrm{L}}\mathbf{W}^{n,l}, \tag{56}$$

$$\begin{aligned} \mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} - \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n}; \mathbf{U}^{\mathrm{H},l}) := & \ \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} \\ & + \sum_{k\in\{1:l-1\}} \tau(a^{\mathsf{i}}_{l,k} - a^{\mathsf{i}}_{l',k})\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n}; \mathbf{U}^{n,k}) \\ & + \sum_{k\in\{1:l-1\}} \tau(a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k})\big(\mathbf{G}^{\mathrm{H}}(\mathbf{U}^{n,k}) - \mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n}; \mathbf{U}^{n,k})\big). \end{aligned} \tag{57}$$

The update $\mathbf{U}^{n,l}$ is obtained by employing the conservative parabolic limiter and by proceeding as for the Euler IMEX scheme. Subtracting (56) from (57) yields

$$\mathbf{U}_i^{\mathrm{H},l} = \mathbf{U}_i^{\mathrm{L},l} + \frac{\tau}{m_i} \sum_{j\in\mathcal{I}(i)} \mathbf{A}_{ij}^{n,l} + \frac{\tau}{m_i}\mathbf{B}_i^{n,l}, \qquad \forall i \in \mathcal{V}, \tag{58}$$

$$\text{with}\quad \mathbf{A}_{ij}^{n,l} := \sum_{k\in\{1:l-1\}} (a_{l,k}^{\mathsf{e}} - a_{l',k}^{\mathsf{e}})\big(\mathbf{D}_{ij}^{\mathrm{H}}(\mathbf{U}^{n,k}) - \mathbf{D}_{ij}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{n,k})\big)$$
$$- (c_l - c_{l'})\mathbf{D}_{ij}^{\mathrm{L,lin}}(\mathbf{W}^{n,l};\mathbf{U}^{\mathrm{L},l}) + \sum_{k\in\{1:l\}} (a_{l,k}^{\mathsf{i}} - a_{l',k}^{\mathsf{i}})\mathbf{D}_{ij}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},l}) \tag{59}$$
$$- \tau^{-1}(m_{ij} - m_i\delta_{ij})\big(\mathbf{U}_j^{\mathrm{H},l} - \mathbf{W}_j^{n,l} - \mathbf{U}_i^{\mathrm{H},l} + \mathbf{W}_i^{n,l}\big),$$

$$\text{and}\quad \mathbf{B}_{i}^{n,l} := \sum_{k\in\{1:l-1\}} (a_{l,k}^{\mathsf{e}} - a_{l',k}^{\mathsf{e}})\big(\mathbf{R}_{i}^{\mathrm{H}}(\mathbf{U}^{n,k}) - \mathbf{R}_{i}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{n,k})\big)$$
$$- (c_l - c_{l'})\mathbf{R}_{i}^{\mathrm{L,lin}}(\mathbf{W}^{n,l};\mathbf{U}^{\mathrm{L},l}) + \sum_{k\in\{1:l\}} (a_{l,k}^{\mathsf{i}} - a_{l',k}^{\mathsf{i}})\mathbf{R}_{i}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},l}). \tag{60}$$

Notice that $\mathbf{A}^{n,l}$ is skew-symmetric and $\mathbf{B}^{n,l} \in \mathcal{B}^I$, in compliance with Definition 2.5. Using the parabolic limiter, we finally set

$$\mathbf{U}^{n,l} := \boldsymbol{\ell}^{\mathrm{par}}(\mathbf{U}^{\mathrm{L},l}, \mathbf{A}^{n,l}, \mathbf{B}^{n,l}). \tag{61}$$

At the end of the loop, the final update is obtained by setting $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1}$.

Remark 3.1 (Quasi-linearization)**.** *We observe that the high-order update* (21) *involves a quasi-linearization based on* $\mathbf{U}^n$*. It is essential that the quasi-linearization for the high-order update be the same for all the stages of the IMEX scheme to preserve the high-order accuracy in time of the method. But, the quasi-linearization for the low-order update at each stage* $l \in \{1{:}s+1\}$ *is based on* $\mathbf{W}^{n,l}$*; this allows us to invoke the invariant-domain property stated in Assumption* (27)*.*

Remark 3.2 (Complexity)**.** *The low-order update* (56) *requires solving a quasi-linear system for all* $l \in \{2{:}s+1\}$*. The high-order update* (57) *requires solving a quasi-linear system for all* $l \in \{2{:}s\}$ *and amounts to an explicit update for* $l = s+1$ *because* $a_{s+1,s+1}^{\mathsf{i}} = 0$*. Thus, the above method requires solving* $(2s-1)$ *quasi-linear systems over each time interval.*

Conclusion

The main result motivating the construction introduced above is the following assertion.

theorem 3.3 (s-stage IDP-IMEX)**.** *Let Assumption* 2.1 *hold and let*

$$\tau \in \Big(1, \frac{\tau^*}{\Delta c^{\max}}\Big]. \tag{62}$$

Assume that the limiters $\boldsymbol{\ell}^{\mathrm{hyp}}$ *and* $\boldsymbol{\ell}^{\mathrm{par}}$ *match Definitions* 2.4-2.5*. Let* $\mathbf{U}^n \in \mathcal{A}^I$*. Consider the* s*-stage IMEX scheme composed of the steps* (51)–(61) *for* $l \in \{2{:}s+1\}$*. This scheme satisfies the following properties:*

(i) *It is well defined;*

(ii) *It is IDP, i.e., it satisfies* (7)*;*

(iii) *It is conservative, i.e., it satisfies* (8)*.*

Proof. Assume (62) and $\mathbf{U}^n \in \mathcal{A}^I$. We are going to show by induction that all the intermediate updates $(\mathbf{U}^{n,l})_{l\in\{1:s+1\}}$ are well defined and are in $\mathcal{A}^I$. The definition $\mathbf{U}^{n,1} := \mathbf{U}^n$ implies that the assumption holds true for $l = 1$. Let now $l \in \{2{:}s+1\}$. We make the following observations: The

low-order hyperbolic update (51) has the same structure as (33); the high-order hyperbolic update (53)-(54) has the same structure as (35)-(36); the low-order parabolic update (56) has the same structure as (38); the high-order parabolic update (58)–(61) has the same structure as (40)–(43). Hence we can apply Lemma 2.7 to the scheme (51)–(61) provided the effective time step $(c_l - c_{l'})\tau$ used in the low-order hyperbolic and parabolic stages (51) and (56) is in the interval $(0, \tau^*]$. But this is the case owing to the assumption (62). Then Lemma 2.7 implies that $\mathbf{U}^{n,l}$ is well defined and is in $\mathcal{A}^I$. This establishes (i) and (ii). Notice that Lemma 2.7 also asserts that

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n,l} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n,l'}, \qquad \forall p \in \mathcal{C}.$$

An induction argument readily gives $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n,l} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_{p,i}^{n}$ for all $l \in \{1{:}s+1\}$, and the conservation property (8) follows from $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1}$. This proves (iii). □

Following Shu and Osher [27] (see also [10, Def. 2.2]), the quantity

$$c_{\text{eff}} := \frac{1}{s\Delta c^{\max}}, \tag{63}$$

is called *efficiency ratio* of the s-stage IMEX scheme. Recall that $\Delta c^{\max}$ is defined in (47). By construction, we have $c_{\text{eff}} \le 1$. Theorem 3.3 shows that the IMEX scheme is IDP for all $\tau \in (0, c_{\text{eff}} s\tau^*]$. Hence it desirable to have an efficiency ratio as large as possible for computational efficiency. In particular, the largest time allowed is $s\times\tau^*$ when $c_{\text{eff}} = 1$. The optimal value $c_{\text{eff}} = 1$ is attained when the coefficients c_j are equi-distributed, i.e., $c_j := \frac{j-1}{s}$ for all $j \in \{1{:}s\}$.

4 Examples of IMEX schemes

In this section, we review some examples of IMEX schemes and introduce some novel schemes. We only consider schemes with p-order accuracy where $p \in \{2, 3, 4\}$. Recall that the IMEX schemes under consideration combine an ERK scheme and an EDIRK scheme. Both schemes consist of $s \ge p$ stages and are described by the Butcher tableaux introduced in (44). In what follows, we use the terminology IMEX$(s, p; c_{\text{eff}})$ for an IMEX scheme with s stages, order p, and efficiency ratio c_{eff}. Four new schemes with optimal efficiency and the following characteristics are introduced in this section:
(1) IMEX$(3, 3; 1)$, singly diagonal, A-stable implicit part, (81);
(2) IMEX$(4, 3; 1)$, singly diagonal, L-stable implicit part, (82);
(3) IMEX$(5, 4; 1)$, singly diagonal, L-stable implicit part, (83);
(4) IMEX$(6, 4; 1)$, singly diagonal, L-stable implicit part, (84).

4.1 Main properties of IMEX schemes

Three important notions for IMEX schemes are the consistency order, the stability of the implicit scheme, and the efficiency ratio. We now briefly discuss these three properties.

For simplicity, we focus on IMEX schemes for which we have $b_i^{\mathsf{e}} = b_i^{\mathsf{i}} =: b_i$ for all $i \in \{1{:}s\}$. We denote by B the row vector in $\mathbb{R}^s$ having components $(b_i)_{i\in\{1:s\}}$. We denote by C the column vector in $\mathbb{R}^s$ having components $(c_j)_{j\in\{1:s\}}$. We also use the notation C^p, $p \ge 0$, for the column vector in $\mathbb{R}^s$ having components $(c_j^p)_{j\in\{1:s\}}$. To be coherent with the literature we set $U := C^0 = (1, \dots, 1)^{\mathsf{T}}$ and use the symbol C instead of C^1. We denote by $B\odot C$ the row vector in $\mathbb{R}^s$ having components $(b_j c_j)_{j\in\{1:s\}}$. We denote by A^{e} (resp., A^{i}) the square matrix of order s with entries $(a_{i,j}^{\mathsf{e}})_{i,j\in\{1:s\}}$ (resp., $(a_{i,j}^{\mathsf{i}})_{i,j\in\{1:s\}}$). Notice that A^{e} is strictly lower triangular, whereas A^{i} is lower triangular.

We adopt the following terminology often used in the literature. The identity matrix of order s is denoted I_s.

Consistency order

Recall that necessary consistency conditions for the explicit and the implicit methods to be separately of order p are

$$BA^{r-1}C^{q-1} = \frac{(q-1)!}{(q-1+r)!}, \qquad \forall r \in \{1{:}p\}, q \in \{1{:}p-r+1\}. \tag{64}$$

These conditions are sufficient for $p \leq 2$. They are also sufficient for all $p \geq 2$ if the ODE systems are autonomous and linear. Additional nonlinear conditions must be enforced for nonlinear autonomous systems when $p \geq 2$. Coupling conditions must be added for IMEX schemes to be of order $p \geq 2$.

The consistency properties of IMEX methods are reviewed in Pareschi and Russo [25, §2.1] and Kennedy and Carpenter [18, §2.2]. The analysis therein is based on the following simplifying assumption (see (45)), which we systematically enforce:

$$A^{\mathsf{e}}U = C, \qquad A^{\mathsf{i}}U = C. \tag{65}$$

The (linear order) conditions to achieve second-order are

$$BU = 1, \qquad BC = \tfrac{1}{2}, \tag{66}$$

while the conditions $BA^{\mathsf{e}}U = BA^{\mathsf{i}}U = \frac{1}{2}$ follow from (65) and (66).

The conditions to achieve third-order accuracy are (65)-(66) together with the following (linear order) conditions

$$BC^2 = \tfrac{1}{3}, \qquad BA^{\mathsf{e}}C = BA^{\mathsf{i}}C = \tfrac{1}{6}, \tag{67}$$

while the conditions $B(A^{\mathsf{e}})^2U = B(A^{\mathsf{i}})^2U = \frac{1}{6}$ follow from (65) and (67).

The conditions to achieve fourth-order accuracy are (65), (66), (67), together with the (linear order) conditions

$$BC^3 = \tfrac{1}{4}, \qquad BA^{\mathsf{e}}C^2 = BA^{\mathsf{i}}C^2 = \tfrac{1}{12}, \qquad B(A^{\mathsf{e}})^2C = B(A^{\mathsf{i}})^2C = \tfrac{1}{24}, \tag{68}$$

the (nonlinear order) condition

$$(B\odot C)A^{\mathsf{e}}C = (B\odot C)A^{\mathsf{i}}C = \tfrac{1}{8}, \tag{69}$$

and the coupling condition

$$BA^{\mathsf{e}}A^{\mathsf{i}}C = BA^{\mathsf{i}}A^{\mathsf{e}}C = \tfrac{1}{24}. \tag{70}$$

The conditions $B(A^{\mathsf{e}})^3U = B(A^{\mathsf{i}})^3U = \frac{1}{24}$ follow from (65) and (68).

Finally, we are also going to make use of the fifth-order linear order conditions for a six-stage, fourth-order method

$$\begin{aligned} &BC^4 = \tfrac{1}{5}, && BA^{\mathsf{e}}C^3 = BA^{\mathsf{i}}C^3 = \tfrac{1}{20}, \\ &B(A^{\mathsf{e}})^2C^2 = B(A^{\mathsf{i}})^2C^2 = \tfrac{1}{60}, && B(A^{\mathsf{e}})^3C = B(A^{\mathsf{i}})^3C = \tfrac{1}{120}. \end{aligned} \tag{71}$$

The conditions $B(A^{\mathsf{e}})^4U = B(A^{\mathsf{i}})^4U = \frac{1}{120}$ follow from (65) and (71).

Stability

The amplification function associated with a DIRK scheme is

$$R(z) := 1 + zB(I_s - zA^{\mathsf{i}})^{-1}U, \qquad z \in \mathbb{C}. \tag{72}$$

Recall that the RK scheme is said to be A-stable if $|R(z)| \leq 1$ for all $z \in \mathbb{C}$ s.t. $\Re(z) \leq 0$ (see Hairer and Wanner [15, Def. IV.3.3]). The scheme is said to be L-stable if it is A-stable and $R(t) \to 0$ as $t \to -\infty$ (see [15, Def. IV.3.7]). For DIRK schemes, A^{i} is invertible if all the diagonal entries of A^{i} are nonzero. In this case, L-stability amounts to $B(A^{\mathsf{i}})^{-1}U = 1$. However, for EDIRK schemes, the first diagonal entry of A^{i} is zero. In this case, one considers the block decompositions

$$A^{\mathsf{i}} = \begin{pmatrix} 0 & 0 \\ \alpha & \tilde{A} \end{pmatrix}, \qquad B = (\beta, \tilde{B}), \tag{73}$$

with $\alpha \in \mathbb{R}^{s-1}$ (column vector), $\tilde{A} \in \mathbb{R}^{s-1,s-1}$, $\beta \in \mathbb{R}$, and $\tilde{B} \in \mathbb{R}^{s-1}$ (row vector). Then, the amplification function defined in (72) can be rewritten as

$$R(z) = 1 + z\beta + z\tilde{B}(I_{s-1} - z\tilde{A})^{-1}(\tilde{U} + z\alpha), \tag{74}$$

where $\tilde{U} \in \mathbb{R}^{s-1}$ is the column vector having all entries equal to one. Assuming that $\tilde{A}$ is invertible, one readily verifies that the EDIRK scheme is L-stable if it is A-stable and if the following holds:

$$\beta = \tilde{B}\tilde{A}^{-1}\alpha, \qquad \tilde{B}\tilde{A}^{-1}\tilde{U} + \tilde{B}\tilde{A}^{-2}\alpha = 1. \tag{75}$$

Notice that the first condition in (75) implies that $\lim_{t\to-\infty} R(t) = 1 - \tilde{B}\tilde{A}^{-1}\tilde{U} - \tilde{B}\tilde{A}^{-2}\alpha$, and the second condition then implies that $\lim_{t\to-\infty} R(t) = 0$.

4.2 Second-order IMEX schemes

A first possibility to obtain a two-stage, second-order IMEX method consists of combining Heun's second-order scheme with the Crank–Nicolson (A-stable) scheme. The Butcher tableaux are

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{76}$$

We have $l'(l) = l - 1$ for all $l \in \{2{:}3\}$, and the efficiency ratio is $c_{\text{eff}} = \frac{1}{2}$. We call this method IMEX$(2, 2; \frac{1}{2})$.

A second possibility consists of combining the explicit and implicit (A-stable) midpoint rules. The corresponding Butcher tableaux are

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 1 & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline 1 & 0 & 1 \end{array} \tag{77}$$

We have $l'(l) = l - 1$ for all $l \in \{2{:}3\}$, and in this case the efficiency ratio reaches the optimal value $c_{\text{eff}} = 1$. We call this method IMEX$(2, 2; 1)$. The amplification function is $R(z) = \frac{2+z}{2-z}$ for the Crank–Nicolson scheme and the midpoint rule. It is remarkable that the amplification function is the same for both schemes. However, the efficiency of the Crank–Nicolson scheme is only $\frac{1}{2}$, whereas that of the midpoint rule is 1.

A third possibility (see Ascher et al. [2, Sec. 2.5]) is to consider a three-stage, second-order scheme in which the implicit scheme is an L-stable, zero-padded, two-stage ESDIRK scheme. The Butcher tableaux are

$$\begin{array}{c|ccc} 0 & 0 & & \\ \gamma & \gamma & 0 & \\ 1 & \delta & 1-\delta & 0 \\ \hline 1 & 0 & 1-\gamma & \gamma \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1 & 0 & 1-\gamma & \gamma \\ \hline 1 & 0 & 1-\gamma & \gamma \end{array} \tag{78}$$

with $\gamma := 1 - \frac{1}{\sqrt{2}} \approx 0.29289$ and δ is an adjustable parameter for which the value $\delta = -\frac{2}{3}\sqrt{2}$ is recommended. We have $l'(l) = l - 1$ for all $l \in \{2{:}4\}$, but the efficiency ratio is only $c_{\text{eff}} = \frac{1}{3}(1-\gamma) \approx 0.24$. We call this method IMEX$(3, 2; 0.24)$.

Remark 4.1 (Strang's splitting)**.** *Strang's splitting can be rewritten as an IMEX scheme. Consider for instance that the explicit (resp., implicit) midpoint rule is used for the explicit (resp., implicit) steps. One can verify that the whole process can be rewritten as a five-stage IMEX scheme with the following Butcher tableaux*

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ \frac{1}{4} & \frac{1}{4} & 0 & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \\ \frac{3}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 \\ \hline 1 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccccc} 0 & 0 & & & & \\ \frac{1}{4} & 0 & 0 & & & \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & & \\ \frac{1}{2} & 0 & 0 & 1 & 0 & \\ \frac{3}{4} & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \end{array}$$

As expected, there is only one implicit substep (the third one). We have $l'(l) = (1, 2, 3, 4, 5)$ for $l \in \{2{:}6\}$ with $\max_{l\in\{2:6\}}(c_l - c_{l'}) = \frac{1}{4}$. Notice that the fourth substep does not involve extra flux computations with respect to the third substep. Hence, the efficiency ratio is $c_{\text{eff}} = 4\tilde{s}^{-1}$ with $\tilde{s} = 4$ (rather than $c_{\text{eff}} = 4s^{-1}$ with $s = 5$), i.e., the method has optimal efficiency. A variant of this method is implemented in [14] to solve the compressible Navier-Stokes equations (the method SPPRK$(3, 3)$ is used therein instead of the midpoint rule though).

4.3 Third-order IMEX schemes

In this section, we consider third-order IMEX schemes composed of three or four stages. Three-stage schemes in which the implicit scheme is A-stable are available in the literature, but none of these methods has optimal efficiency. We derive here a three-stage IMEX scheme achieving optimality and whose implicit tableau is A-stable. We also construct a four-stage scheme with optimal efficiency whose implicit tableau is L-stable.

4.3.1 Three-stage schemes

A first possibility to obtain a three-stage, third-order IMEX method consists of using the two-stage, third-order, zero-padded ESDIRK scheme (Crouzeix [8], Nørsett [22]) for the implicit scheme and combining it with the three-stage, third-order ERK scheme sharing the same coefficients c_j and b_j. The corresponding Butcher tableaux are (see Ascher et al. [2, Sec. 2.4])

$$\begin{array}{c|ccc} 0 & 0 & & \\ \gamma & \gamma & 0 & \\ 1-\gamma & \gamma-1 & 2-2\gamma & 0 \\ \hline 1 & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1-\gamma & 0 & 1-2\gamma & \gamma \\ \hline 1 & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \tag{79}$$

with $\gamma := \frac{1}{2} + \frac{1}{2\sqrt{3}} \approx 0.78867$ (i.e., $\gamma^2 = \gamma - \frac{1}{6}$). The amplification function is

$$R(z) = \frac{1 + (1-2\gamma)z + (\frac{1}{3} - \gamma)z^2}{(1-\gamma z)^2}. \tag{80}$$

The zero-padded ESDIRK scheme is A-stable, but not L-stable because we only have $\lim_{t\to-\infty} R(t) = 1 - \sqrt{3} \approx -0.73205$. Finally, we observe that the values for l' are $(1, 1, 2)$, and the efficiency ratio is only $c_{\text{eff}} = \frac{1}{3}\gamma \approx 0.26$. We call this method IMEX$(3, 3; 0.26)$.

We now propose a three-stage, third-order IMEX method with optimal efficiency. We call this method IMEX$(3, 3; 1)$. We use the third-order Heun method for the ERK part, and we design the corresponding three-stage, third-order EDIRK scheme. To this purpose, we first request that the EDIRK scheme has the same set of coefficients c_j and b_j as Heun's method, so that there remains to determine the matrix A^{i}. Since this matrix is lower triangular, of order three, and $a^{\text{i}}_{1,1} = 0$, this leaves five entries to be determined. Four equations can be enforced: two from Butcher's simplifying assumption (65) (there are two equations corresponding to the rows $i \in \{2, 3\}$ in (65) since the row corresponding to $i = 1$ is trivial), one is the (linear order) condition $BA^{\text{i}}C = \frac{1}{6}$ stated in (67) (the remaining linear order conditions are already satisfied), and one is the first stability condition in (75). One can show that it is not possible to enforce the second equality in (75) (there would be no solution). The fifth condition we use to close the system consists of minimizing $\lim_{t\to-\infty} R(t)$. Solving this problem leads to an A-stable method with $\lim_{t\to-\infty} R(t) = 1 - \sqrt{3}$. Incidentally, the implicit scheme turns out to be singly diagonal, although this property has not been enforced explicitly. The Butcher arrays of the ERK and ESDIRK scheme are as follows:

$$\begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3} & 0 & \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3} - \gamma & \gamma & \\ \frac{2}{3} & \gamma & \frac{2}{3} - 2\gamma & \gamma \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \tag{81}$$

with (again) $\gamma := \frac{1}{2} + \frac{1}{2\sqrt{3}} \approx 0.78867$. We have $l'(l) = l - 1$ for all $l \in \{2{:}4\}$, and the efficiency ratio reaches the optimal value $c_{\text{eff}} = 1$. Quite remarkably, the amplification function for the above ESDIRK scheme is still given by (80). Hence, the amplification functions of the methods described by the Butcher tableaux (79) and (81) are identical, but the efficiency of (79) is only $\frac{1}{3}\gamma \approx 0.26$, whereas that of the new method (81) is 1.

4.3.2 Four-stage schemes

It is possible to devise a four-stage, third-order IMEX method with optimal efficiency in which the implicit part is an ESDIRK L-stable scheme. We call this method IMEX$(4, 3; 1)$. We set $c_l := \frac{l-1}{4}$ for all $l \in \{1{:}4\}$ to achieve optimal efficiency. There are 13 coefficients to be determined: 9 entries in the matrix A^{i} and the four components of the vector b. We enforce Butcher's simplifying assumption (65) (3 equations), the (linear) order conditions (66) and (67) (4 equations), and the two conditions in (75) which are necessary for L-stability. This gives 9 equations. We additionally require that the scheme be singly diagonal, giving the two additional equations $a^{\text{i}}_{2,2} = a^{\text{i}}_{3,3} = a^{\text{i}}_{4,4}$, and that $BC^3 = \frac{1}{4}$ (this is the first of the fourth-order (linear) conditions in (68)). The resulting under-determined set of nonlinear equations (12 equations, 13 unknowns) is solved using `julia` with 10^{-15} tolerance. The following solution is found:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{4} & -0.1858665215084591 & 0.4358665215084591 & & \\ \frac{1}{2} & -0.4367256409878701 & 0.5008591194794110 & 0.4358665215084591 & \\ \frac{3}{4} & -0.0423391342724147 & 0.7701152303135821 & -0.4136426175496265 & 0.4358665215084591 \\ \hline 1 & 0 & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{array} \tag{82a}$$

This ESDIRK scheme is L-stable.

The companion ERK method that shares the same set of coefficients c_j and b_j has already been proposed in Ern and Guermond [10]. The six coefficients of the matrix A^{e} are obtained by enforcing the fourth-order linear consistency conditions (three linear equations from (65), one linear equation form (67), one linear and one nonlinear equation from (68)) This is the only four-stage, third-order ERK method with optimally distributed coefficients that is also fourth-order accurate on linear problems. Its Butcher tableau is as follows:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{4} & \frac{1}{4} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ \frac{3}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 \\ \hline 1 & 0 & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{array} \tag{82b}$$

Near the origin along the imaginary axis, we have $|R^{\mathsf{e}}(i\epsilon)| = 1+\rho_6^{\mathsf{e}}\epsilon^6+\mathcal{O}(\epsilon^8)$ with $\rho_6^{\mathsf{e}} = -2B(A^{\mathsf{e}})^4C+2B(A^{\mathsf{e}})^3C - B(A^{\mathsf{e}})^2C + \frac{1}{36} = -\frac{1}{72}$.

Remark 4.2 (Other four- and five-stage methods)**.** *A third-order IMEX method combining a four-stage ERK method with a four-stage L-stable DIRK method is described in Ascher et al. [2, Sec. 2.7] (the DIRK method has actually three stages since the third and fourth stages are identical). The efficiency ratio is close to* $c_{\text{eff}} = 0.46$*. A variant of the implicit four-stage scheme is studied in Calvo et al. [6]. A method combining a five-stage ERK method with a five-stage L-stable DIRK method is described in [2, Sec. 2.8] (the DIRK method has actually four stages since the fourth and fifth stages are identical). The efficiency ratio is only* $c_{\text{eff}} = \frac{1}{8}$*.*

4.4 Fourth-order IMEX schemes

Some fourth-order IMEX methods composed of five and six stages are discussed in Kennedy and Carpenter [18, Sec. 3.2 & 3.3], but the efficiency ratio of these methods is far from being optimal. Here, we devise five- and six-stage, fourth-order IMEX schemes with optimal efficiency; the implicit scheme is L-stable in both cases. We call these methods IMEX$(5,4;1)$ and IMEX$(6,4;1)$.

4.4.1 Five-stage scheme

We set $c_l := \frac{l-1}{5}$ for all $l \in \{1{:}5\}$ to achieve optimal efficiency. There are 19 coefficients to be determined: 14 entries in the matrix A^{i} and the five components of the vector b. We enforce Butcher's simplifying assumption (65) (4 equations), the (linear) order conditions (66), (67), and (68) (7 equations), the (nonlinear) order condition (69) (one equation), and the two conditions in (75) which are necessary for L-stability. This gives 14 equations. We additionally require that the scheme be singly diagonal, giving the three additional equations $a^{\mathsf{i}}_{2,2} = a^{\mathsf{i}}_{3,3} = a^{\mathsf{i}}_{4,4} = a^{\mathsf{i}}_{5,5}$, and that $a^{\mathsf{i}}_{5,1} = 0$, giving one additional equation. The resulting under-determined set of nonlinear equations (18 equations, 19 unknowns) is solved using `julia` with 2×10^{-16} tolerance. The following solution is found:

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ \frac{1}{5} & -0.37281606248213511 & 0.57281606248213512 & & & \\ \frac{2}{5} & -0.66007935107985416 & 0.48726328859771911 & 0.57281606248213512 & & \\ \frac{3}{5} & -0.69934543274239502 & 1.82596107935553742 & -1.09943170909527743 & 0.57281606248213512 & \\ \frac{4}{5} & 0 & -0.05144383172900784 & 1.17898889035791732 & -0.90036112111104449 & \dots \\ \hline 1 & -0.10511678454691901 & 0.87880047152100838 & -0.58903404061484477 & 0.46213380485434047 & \dots \end{array}$$

$$\begin{array}{c|cc} \frac{4}{5} & \dots & 0.57281606248213512 \\ \hline 1 & \dots & 0.35321654878641495 \end{array} \tag{83a}$$

This ESDIRK scheme is L-stable. Along the imaginary axis near the origin, we have $|R^{\mathsf{i}}(\mathrm{i}\epsilon)| = 1 + \rho_6^{\mathsf{i}}\epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho_6^{\mathsf{i}} \approx -0.0846$, where $\rho_6^{\mathsf{i}} = -2B(A^{\mathsf{i}})^4C + 2B(A^{\mathsf{i}})^3C - \frac{1}{72}$.

We now devise the companion ERK scheme that shares the same set of coefficients c_j and b_j. There are 10 unknowns (the entries of the strictly lower triangular matrix A^{e}). We enforce Butcher's simplifying assumption (65) (4 equations) and the (linear) order conditions (67) involving the matrix A^{e} (three equations, since the remaining order conditions have already been accounted for in the design of the ESDIRK scheme above), the (nonlinear) order condition (69) (one equation), and the two coupling conditions (70). This gives 10 equations. The resulting set of nonlinear equations (10 equations, 10 unknowns) is solved using `julia` with 2×10^{-16} tolerance. The following solution is found:

$$
\begin{array}{c|ccccc}
0 & 0 & & & & \\
\frac{1}{5} & 0.2 & 0 & & & \\
\frac{2}{5} & 0.26075582269554909 & 0.13924417730445096 & 0 & & \\
\frac{3}{5} & -0.25856517872570289 & 0.91136274166280729 & -0.05279756293710430 & 0 & \\
\frac{4}{5} & 0.21623276431503774 & 0.51534223099602405 & -0.81662794199265554 & 0.88505294668159373 & \dots \\
\hline
1 & -0.10511678454691901 & 0.87880047152100838 & -0.58903404061484477 & 0.46213380485434047 & \dots
\end{array}
\qquad
\begin{array}{c|cc}
\frac{4}{5} & \dots & 0 \\
\hline
1 & \dots & 0.35321654878641495
\end{array}
\tag{83b}
$$

We have $|R^{\mathsf{e}}(\mathrm{i}\epsilon)| = 1 + \rho_6^{\mathsf{e}}\epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho_6^{\mathsf{e}} \approx -0.0148$ along the imaginary axis near the origin, where $\rho_6^{\mathsf{e}} = 2B(A^{\mathsf{e}})^3C - \frac{1}{72}$ (notice that $2B(A^{\mathsf{e}})^4C = 0$).

4.4.2 Six-stage schemes

A six-stage, fourth-order method with L-stable (actually, stiffly accurate) implicit scheme is designed in Calvo et al. [6] using a L-stable, six-stage (actually five distinct stages) fourth-order SDIRK method from Hairer and Wanner [15]. More precisely, the explicit tableau is given by Equation (14) in [6] and the implicit tableau is given by Equation (IV.6.16) in [15] (see also Table IV.6.5). The efficiency ratio of this method is only $c_{\text{eff}} = \frac{1}{12} \approx 0.08$. We call this method IMEX$(6, 4; 0.08)$.

We now propose a six-stage, fourth-order IMEX method with optimal efficiency. We set $c_l := \frac{l-1}{6}$ and $l'(l) = l - 1$ for all $l \in \{1{:}6\}$ to achieve optimal efficiency. There are 26 coefficients to be determined for the EDIRK scheme: 20 entries in the matrix A^{i} and the six components of the vector b. We enforce Butcher's simplifying assumption (65) (5 equations), the (linear) order conditions (66), (67), (68) (7 equations), the (nonlinear) order condition (69) (one equation), and the two conditions in (75) which are necessary for L-stability. We also enforce the fifth-order linear order conditions (71) (4 equations). This gives 19 equations. We additionally require that the scheme be singly diagonal, giving the four equations $a^{\mathsf{i}}_{2,2} = a^{\mathsf{i}}_{3,3} = a^{\mathsf{i}}_{4,4} = a^{\mathsf{i}}_{5,5} = a^{\mathsf{i}}_{6,6}$. We also set $b_4 = 0.47$, giving six additional equations. The resulting set of nonlinear equations (24 equations, 26 unknowns) is solved using `julia` with 4×10^{-16} tolerance. The following solution is found:

$$
\begin{array}{c|ccccc}
0 & 0 & & & & \\
\frac{1}{6} & -0.1113871744697862 & 0.2780538411364528 & & & \\
\frac{2}{6} & -0.7193507615705692 & 0.7746302537674498 & 0.2780538411364528 & & \\
\frac{3}{6} & 0.5518029866688972 & 0.1104050865166429 & -0.4402619143219927 & 0.2780538411364528 & \\
\frac{4}{6} & 0.2044212940947437 & 0.7369116313032833 & -0.6137248254193539 & 0.0610047255515406 & \dots \\
\frac{5}{6} & 0.0660767687645300 & 0.0489052670268613 & 0.2501367454670004 & 0.5829521002593755 & \dots \\
\hline
1 & 0.083 & 0.135 & 0.13 & 0.47 & \dots
\end{array}
\qquad
\begin{array}{c|cc}
\frac{4}{6} & \dots & 0.2780538411364528 \\
\frac{5}{6} & \dots & -0.3927913893208868 \quad 0.2780538411364528 \\
\hline
1 & \dots & -0.285 \qquad\qquad 0.467
\end{array}
\tag{84a}
$$

We have $|R^{\mathsf{i}}(\mathrm{i}\epsilon)| = 1+\rho^{\mathsf{i}}_6\epsilon^6+\mathcal{O}(\epsilon^8)$ along the imaginary axis near the origin, with $\rho^{\mathsf{i}}_6 \approx -1.06\times10^{-3}$, where we recall that $\rho^{\mathsf{i}}_6 := -2B(A^{\mathsf{i}})^4C + 2B(A^{\mathsf{i}})^3C - \frac{1}{72} = -2B(A^{\mathsf{i}})^4C + \frac{1}{360}$.

We now proceed to find a companion ERK scheme sharing the same set of coefficients c_j and b_j. There are 15 unknowns (the entries of the strictly lower triangular matrix A^{e}). We enforce Butcher's simplifying assumption (65) (5 equations), the (linear) order conditions (67), (68) involving the matrix A^{e} (three equations), the (nonlinear) order condition (69) (one equation), and the two coupling conditions (70). This gives 11 equations. We also enforce the fifth-order linear order conditions (71) (4 equations). In total we have 14 equations. The resulting under-determined set of nonlinear equations (14 equations, 15 unknowns) is solved using `julia`. The following solution is found with 4×10^{-16} tolerance:

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & \\
\frac{1}{6} & 0.1666666666666667 & 0 & 0 & 0 & \\
\frac{2}{6} & -0.4447518666865896 & 0.7780852000199229 & 0 & 0 & \\
\frac{3}{6} & 0.0893971199002357 & 0.1913734465774906 & 0.2192294335222737 & 0 & \\
\frac{4}{6} & 0.0635170175925033 & 0.1428758587504802 & 0.1359933602040186 & 0.3242804301196646 & \\
\frac{5}{6} & 0.0727304753901258 & 0.2698992458411843 & -0.0619049508228351 & 0.2187862524098492 & \dots \\
\hline
1 & 0.083 & 0.135 & 0.13 & 0.47 & \dots
\end{array}
\qquad
\begin{array}{c|ccc}
\frac{5}{6} & \dots & 0.3338223105150092 & 0 \\
\hline
1 & \dots & -0.285 & 0.467
\end{array}
\tag{84b}
$$

We have $|R^{\mathsf{e}}(\mathrm{i}\epsilon)| = 1+\rho^{\mathsf{e}}_6\epsilon^6+\mathcal{O}(\epsilon^8)$ along the imaginary axis near the origin, with $\rho^{\mathsf{e}}_6 \approx -9.67\times10^{-5}$, where $\rho^{\mathsf{e}}_6 := -2B(A^{\mathsf{e}})^4C + 2B(A^{\mathsf{e}})^3C - \frac{1}{72} = -2B(A^{\mathsf{e}})^4C + \frac{1}{360}$. The value of the coefficient b_4 is adjusted to the value 0.47 to make ρ^{e}_6 negative.

5 Numerical illustrations

We illustrate the IMEX methods proposed in the paper. We start with convergence tests on an ODE system. Then we solve a scalar nonlinear conservation equation with hyperbolic and parabolic fluxes.

5.1 Convergence tests

We test the convergence properties of the new IMEX methods proposed in this paper and compare them to the other published methods listed in Section 4. Following Kennedy and Carpenter [18, §5.1], we consider the 2×2 ODE system

$$\partial_t y_1(t) = -2y_1 + \epsilon^{-1}(y_2^2 - y_1), \qquad \partial_t y_2(t) = y_1 - y_2 - y_2^2, \tag{85}$$

with $\epsilon > 0$ and initial condition $y_1(0) = y_2(0) = 1$. The solution is $y_1(t) = y_2^2(t)$, $y_1(t) = e^{-t}$. As $\epsilon \to 0$, the above problem degenerates into the index-1 differential algebraic equation $\partial_t y_2(t) = y_1 - y_2 - y_2^2$ with $y_1 = y_2^2$. We denote by $\mathbf{U} := (\mathsf{u}_1, \mathsf{u}_2)^{\mathsf{T}}$ the approximate solution produced by the IMEX methods. Referring to (11) for the notation, we set $\mathbb{M} := \mathbb{I}_2$, where $\mathbb{I}_2$ is the 2×2 identity matrix, and

$$\mathbf{F}(\mathbf{U}) := (-2\mathsf{u}_1, \mathsf{u}_1 - \mathsf{u}_2 - \mathsf{u}_2^2)^{\mathsf{T}}, \quad \mathbf{G}(\mathbf{U}) := (\epsilon^{-1}(\mathsf{u}_2^2 - \mathsf{u}_1), 0)^{\mathsf{T}}, \quad \mathbf{G}^{\mathrm{lin}}(\mathbf{W}, \mathbf{U}) := \mathbf{G}(\mathbf{U}). \tag{86}$$

Notice that $\mathbf{G}^{\mathrm{lin}}$ is not linear, but solving the problem (10) is simple:

$$(\mathbb{I} - \tau\mathbf{G}^{\mathrm{lin}}(\mathbf{W}, \cdot))^{-1}(\mathbf{W}) = (\tfrac{1}{\epsilon+\tau}(\epsilon\mathsf{w}_1 + \tau\mathsf{w}_2^2), \mathsf{w}_2)^{\mathsf{T}}. \tag{87}$$

We test all the methods mentioned in Section 4 by solving the above problem over the time interval $[0, T]$ with $T = 4$, the initial data $(\mathsf{u}_1(0), \mathsf{u}_2(0)) = (1, 1)$, and for $\epsilon \in \{1, 10^{-6}\}$. For each method, we compute the errors $|\mathsf{u}_1(T) - y_1(T)|/|y_1(T) + y_2(T)|$ and $|\mathsf{u}_2(T) - y_2(T)|/|y_1(T) + y_2(T)|$. The results are reported in Figures 1, 2, 3. The symbols "y1,e0", "y1,e-6", "y2,e0", and "y2,e-6" in the legend refer to the error on the variable y_1 with $\epsilon = 10^0$, the error on the variable y_1 with $\epsilon = 10^{-6}$, the error on the variable y_2 with $\epsilon = 1$, and the error on the variable y_2 with $\epsilon = 10^{-6}$, respectively.

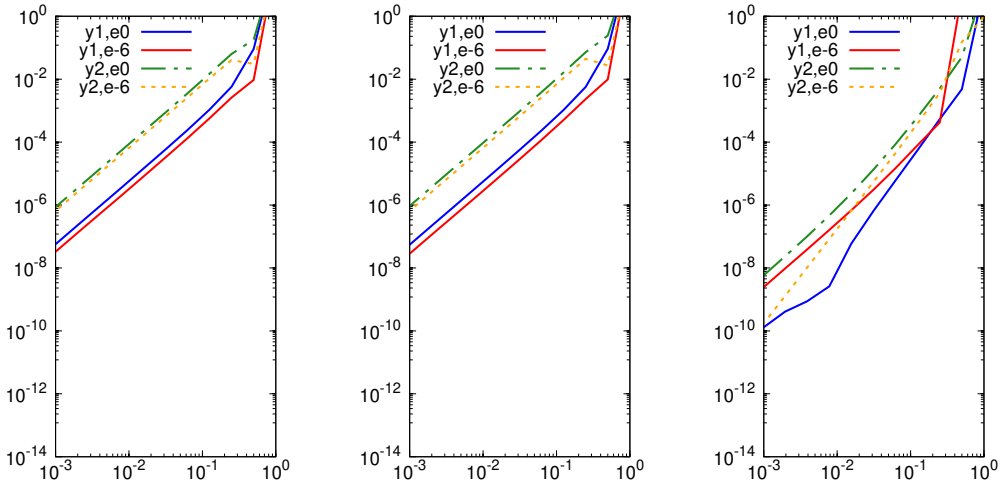

Figure 1: Convergence test on problem (85) for the second-order methods IMEX$(2, 2; \frac{1}{2})$, IMEX$(2, 2; 1)$, IMEX$(3, 2; 0.24)$ (with the Butcher tableaux (76), (77), (78) from left to right).

The three panels in Figure 1 show the results for the second-order methods IMEX$(2, 2; \frac{1}{2})$, IMEX$(2, 2; 1)$ and IMEX$(3, 2; 0.24)$ (from left to right and with the Butcher tableaux (76), (77), (78)). We observe that they all deliver second-order accuracy uniformly with respect to ϵ for both y_1 and y_2.

The results for the methods IMEX$(3, 3; 0.26)$, IMEX$(3, 3; 1)$, IMEX$(4, 3; 1)$ (see (79), (81), and (82) respectively) are shown in Figure 2. We observe third-order accuracy for the three methods on both variables when $\epsilon = 1$, but as expected, the convergence on y_1 reduces to second-order when $\epsilon = 10^{-6}$. This order reduction in the pre-asymptotic range (i.e., $\epsilon < \tau$) is well documented in the literature and we refer the reader to Boscarino and Pareschi [4] for an analysis of this phenomenon. We observe that the two new methods introduced in this paper (i.e., IMEX$(3, 3; 1)$, IMEX$(4, 3; 1)$) perform as expected. Recall that IMEX$(3, 3; 0.26)$ and IMEX$(3, 3; 1)$ are both three-stage, third-oder methods. The efficiency of IMEX$(3, 3; 0.26)$ is only 0.26 whereas that of IMEX$(3, 3; 1)$ is optimal. The implicit components of these methods are A-stable. They also have the same amplification functions. The method IMEX$(4, 3; 1)$ is composed of four stages, has efficiency 1, and its implicit component is L-stable.

Finally, the results for the three fourth-order methods IMEX$(5; 4; 1)$ (see (83)), IMEX$(6; 4; 0.08)$, and IMEX$(6, 4; 1)$ (see (84)) are shown in Figure 3. We observe fourth-order accuracy for the three methods on both variables when $\epsilon = 1$. The convergence rate on y_2 is still 4 when $\epsilon = 10^{-6}$, but the convergence rate for y_1 reduces to second order when $\epsilon = 10^{-6}$. We observe that the two new methods introduced in this paper (i.e., IMEX$(5; 4; 1)$ and IMEX$(6; 4; 1)$) perform as expected. Recall that IMEX$(5; 4; 1)$ has five stages with efficiency $c_{\text{eff}} = 1$, IMEX$(6; 4; 0.08)$ has six stages with efficiency $c_{\text{eff}} = 0.08$, and IMEX$(6; 4; 1)$ has six stages with efficiency $c_{\text{eff}} = 1$. The implicit

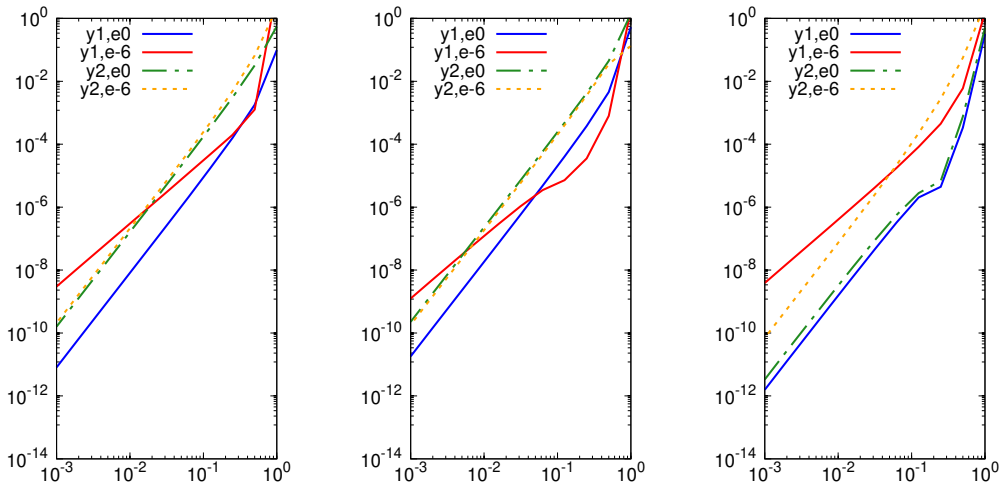

Figure 2: Convergence test on problem (85) for the third-order methods IMEX$(3,3;0.26)$, IMEX$(3,3;1)$; IMEX$(4,3;1)$ (with the Butcher tableaux (79), (81), (82) from left to right).

components of these three methods are L-stable.

5.2 Nonlinear scalar conservation equation

In this section, we illustrate the method on the following scalar nonlinear conservation equation:

$$\partial_t u + \nabla\cdot\boldsymbol{f}(u) - \epsilon\Delta u = 0, \quad \boldsymbol{x}\in D_\infty,\ t>0, \tag{88}$$

posed in the two-dimensional domain $D_\infty := \mathbb{R}\times(0,1)$. The flux is defined by $\boldsymbol{f}(u) := (u(1-u),0)^\mathsf{T}$. With the notation $\boldsymbol{x} := (x,y)$, the initial data is

$$u_0(\boldsymbol{x}) := \mu + \delta\tanh\left(\tfrac{\delta}{\epsilon}(x-x_0)\right), \quad \mu := \tfrac{1}{2}(u_L+u_R), \quad \delta := \tfrac{1}{2}(u_R-u_L). \tag{89}$$

Assuming homogeneous Neuman boundary conditions on the top and bottom parts of the domain, the solution to this Cauchy problem is a wave moving at speed $s := 1-2\mu$:

$$v(\boldsymbol{x},t) = u_0(\boldsymbol{x}-\boldsymbol{s}t) \quad \text{with} \quad \boldsymbol{s} := (s,0). \tag{90}$$

The method described in this paper is implemented using continuous finite elements. The tests are done with continuous $\mathbb{P}_1$ and $\mathbb{P}_3$ finite elements. The low-order solution method for the hyperbolic subproblem is fully described in [11]. The high-order method and the limiting are described in [12, 13]. We use FCT to perform the limiting as the problem is scalar-valued. Local bounds are used at every grid point. Relaxation of the bounds guaranteeing high-order convergence is done as explained in [12, §4.7.1] and [13, §7.6].

We set $u_L := -1$ and $u_R := 1$, so that the solution to (88) is a wave moving at speed $s = 1$. The numerical simulations are done in the truncated computational domain $D := (x_L,x_R)\times(y_B,y_T)$ with $x_L = y_B := 0$, $x_R := 1$, $y_T := \frac{1}{4}$. Let $\partial D_{\mathrm{D}}^{\mathrm{hyp}} = \partial D_{\mathrm{D}}^{\mathrm{par}} := \{x_L,x_R\}\times(y_B,y_T)$, $:= (\{x_L\}\cup\{x_R\})\times(y_B,y_T)$, and $\partial D_{\mathrm{N}}^{\mathrm{par}} := (x_L,x_R)\times\{y_B,y_T\}$ (it happens here that $\partial D_{\mathrm{D}}^{\mathrm{hyp}} = \partial D_{\mathrm{D}}^{\mathrm{par}}$ because $1-2u_R<0$). At each stage of the IMEX method, Dirichlet boundary conditions are enforced at $\partial D_{\mathrm{D}}^{\mathrm{hyp}}$ for the hyperbolic subproblems and at $\partial D_{\mathrm{D}}^{\mathrm{par}}$ for the parabolic subproblems. Homogeneous

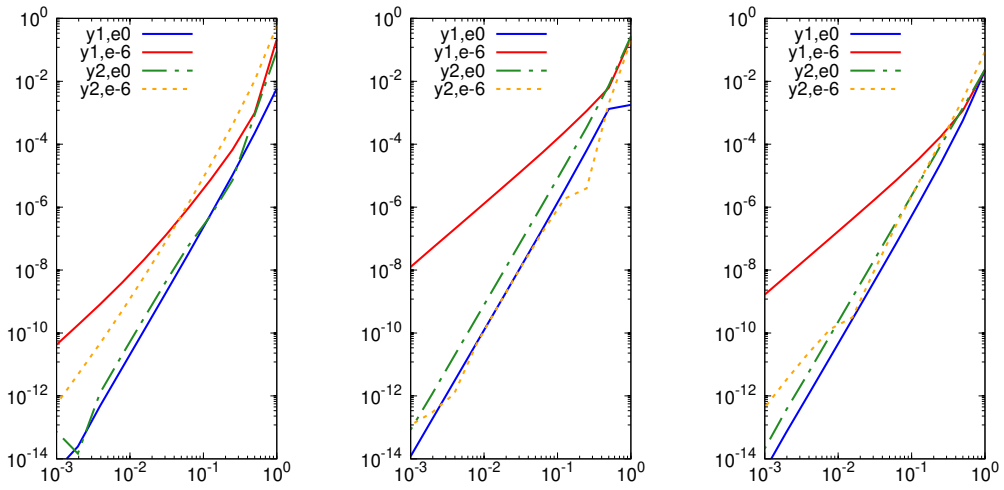

Figure 3: Convergence test on problem (85) for the fourth-order methods IMEX$(5;4;1)$, IMEX$(6;4;0.08)$ and IMEX$(6,4;1)$ (with the Butcher tableaux (83) and (84) for IMEX$(5;4;1)$ and IMEX$(6,4;1)$). IMEX$(6;4;0.08)$ is defined in the first paragraph of §4.4.2.

Neumann conditions are enforced on $\partial D_{\mathrm{N}}^{\mathrm{par}}$ for the parabolic subproblems. The enforcement of the boundary condition for the hyperbolic subproblems is done at the end of each stage of the IMEX step.

In all the tests, the time step is computed by using the expression

$$\tau := \mathrm{CFL}\times s\times\tau^*, \tag{91}$$

where $\mathrm{CFL} > 0$ is a fixed parameter, s is the number of stages of the IMEX method, and τ^* is the maximum time step for which the low-order hyperbolic update is IDP; see Assumption 2.1(i). This definition guarantees that for a given simulation time T, the total number of flux evaluations (which is a measure of the algorithmic cost) is approximately $s\times\frac{T}{\tau} = s\times\frac{T}{\mathrm{CFL}\times s\times\tau^*} = \frac{T}{\mathrm{CFL}\times\tau^*}$. Hence, for a given mesh, a given final time T, and a given CFL number, the algorithmic cost of two IMEX methods with different number of stages is approximately identical. The simulations are done up to $T = \frac{1}{2}$ for $\epsilon = 2\times10^{-n}$, $n \in \{2,3,4\}$. We use unstructured Delaunay meshes. We test the following five methods: IMEX$(2,2;1)$; IMEX$(3,3;1)$; IMEX$(4,3;1)$; IMEX$(5,4;1)$; and IMEX$(6,4;1)$. All the errors are evaluated at T and are relative.

We show the errors and the convergence rates for continuous $\mathbb{P}_1$ elements in Table 5.2. We observe that the methods deliver second-order accuracy when the mesh size is small enough to capture the viscous layer of size ϵ. The accuracy is limited to second-order due to our using $\mathbb{P}_1$ elements. We also notice that all the methods deliver first-order accuracy when the mesh size cannot capture the viscous layer. First-order accuracy is optimal in this case.

We show the errors and the convergence rates for continuous $\mathbb{P}_3$ elements in Table 5.2. The methods deliver optimal accuracy when the mesh size is small enough to capture the viscous layer, that is, second-order for IMEX$(2,2;1)$ and fourth-order for the other methods. There seems to be some super-convergence effect for the third-order methods IMEX$(3,3;1)$ and IMEX$(4;4;1)$. Here again, all the methods deliver first-order accuracy when the mesh size cannot capture the viscous layer.

Table 1: Problem (88) for $\epsilon = 2\times10^{-n}$, $n \in \{2,3,4\}$. $\mathbb{P}_1$ finite elements. Error in the L^1-norm. First row: IMEX$(2,2;1)$. Second row: IMEX$(3,3;1)$ and IMEX$(4,3;1)$. Third row IMEX$(5,4;1)$ and IMEX$(6,4;1)$.

	$\epsilon=10^{-2}$		$\epsilon=10^{-3}$		$\epsilon=10^{-4}$	
I	IMEX(2,1;1)	rate	(2,1;1)	rate	(2,1;1)	rate
106	1.98E-02	–	3.36E-02	–	3.60E-02	–
360	4.13E-03	2.56	1.61E-02	1.20	1.52E-02	1.41
1309	8.12E-04	2.52	7.60E-03	1.16	7.64E-03	1.07
4825	2.03E-04	2.13	2.88E-03	1.49	4.02E-03	0.98
18846	4.99E-05	2.06	7.01E-04	2.07	1.99E-03	1.03
74510	1.25E-05	2.01	1.29E-04	2.46	9.83E-04	1.03

	$\epsilon=10^{-2}$				$\epsilon=10^{-3}$				$\epsilon=10^{-4}$			
I	IMEX(3,3;1)	rate	IMEX(4,3;1)	rate	(3,3;1)	rate	(4,3;1)	rate	(3,3;1)	rate	(4,3;1)	rate
106	1.97E-02	–	1.97E-02	–	3.36E-02	–	3.36E-02	–	3.60E-02	–	3.60E-02	–
360	4.13E-03	2.56	4.13E-03	2.56	1.62E-02	1.20	1.61E-02	1.20	1.52E-02	1.41	1.52E-02	1.41
1309	8.26E-04	2.49	8.27E-04	2.49	7.62E-03	1.17	7.60E-03	1.17	7.66E-03	1.06	7.63E-03	1.06
4825	2.04E-04	2.15	2.04E-04	2.15	2.88E-03	1.49	2.88E-03	1.49	4.03E-03	0.98	4.01E-03	0.98
18846	5.00E-05	2.06	5.00E-05	2.06	7.03E-04	2.07	7.04E-04	2.07	2.00E-03	1.03	1.99E-03	1.03
74510	1.25E-05	2.02	1.25E-05	2.02	1.32E-04	2.44	1.32E-04	2.44	9.84E-04	1.03	9.82E-04	1.03

	$\epsilon=10^{-2}$				$\epsilon=10^{-3}$				$\epsilon=10^{-4}$			
I	IMEX(5,4;1)	rate	IMEX(6,4;1)	rate	(5,4;1)	rate	(6,4;1)	rate	(5,4;1)	rate	(6,4;1)	rate
106	1.98E-02	–	1.97E-02	–	3.36E-02	–	3.35E-02	–	3.60E-02	–	3.59E-02	–
360	4.13E-03	2.56	4.10E-03	2.57	1.62E-02	1.20	1.59E-02	1.22	1.52E-02	1.41	1.50E-02	1.42
1309	8.26E-04	2.49	8.11E-04	2.51	7.65E-03	1.16	7.42E-03	1.18	7.65E-03	1.07	7.56E-03	1.06
4825	2.03E-04	2.15	2.03E-04	2.13	2.93E-03	1.47	2.81E-03	1.49	4.01E-03	0.99	3.98E-03	0.99
18846	4.99E-05	2.06	4.99E-05	2.06	7.22E-04	2.06	6.95E-04	2.05	1.99E-03	1.03	1.96E-03	1.04
74510	1.25E-05	2.02	1.25E-05	2.02	1.33E-04	2.47	1.28E-04	2.46	9.86E-04	1.02	9.63E-04	1.04

6 Conclusions

A new time stepping technique making every IMEX method invariant-domain preserving has been introduced. New IMEX methods with optimal efficiency have been constructed. The numerical experiments done in §5.1 and §5.2 demonstrate that the new IMEX methods proposed in this paper behave as predicted by the theory. All the methods tested are invariant-domain preserving and deliver the expected accuracy. A natural perspective of this work is to show how present methodology can be used to solve nonlinear systems of conservation equations like the compressible Navier–Stokes equations.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation via grants DMS2110868; the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397; by the Army Research Office under grant/contract number W911NF-19-1-0431; and the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contracts B640889. The support of INRIA through the International Chair program is also acknowledged.

References

[1] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton. Implicit-explicit methods for time-dependent partial differential equations. *SIAM J. Numer. Anal.*, 32(3):797–823, 1995.

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997.

Table 2: Problem (88) for $\epsilon = 2{\times}10^{-n}$, $n \in \{2,3,4\}$. $\mathbb{P}_3$ finite elements. Error in the L^1-norm. First row: IMEX$(2,2;1)$. Second row: IMEX$(3,3;1)$ and IMEX$(4,3;1)$. Third row IMEX$(5,4;1)$ and IMEX$(6,4;1)$.

	$\epsilon = 10^{-2}$		$\epsilon = 10^{-3}$		$\epsilon = 10^{-4}$	
I	IMEX (2,1;1)	rate	(2,1;1)	rate	(2,1;1)	rate
778	1.48E-03	–	1.80E-02	–	1.94E-02	–
2902	2.10E-05	6.46	6.79E-03	1.48	8.16E-03	1.31
11113	1.18E-06	4.29	2.07E-03	1.77	4.04E-03	1.05
42097	1.45E-07	3.14	2.95E-04	2.92	2.00E-03	1.06
166966	3.02E-08	2.28	6.82E-06	5.47	9.09E-04	1.14
665302	7.44E-09	2.03	2.62E-07	4.72	3.20E-04	1.51

	$\epsilon = 10^{-2}$				$\epsilon = 10^{-3}$				$\epsilon = 10^{-4}$			
I	IMEX (3,3;1)	rate	IMEX (4,3;1)	rate	(3,3;1)	rate	(4,3;1)	rate	(3,3;1)	rate	(4,3;1)	rate
778	1.48E-03	–	1.48E-03	–	1.80E-02	–	1.80E-02	–	1.93E-02	–	1.93E-02	–
2902	2.01E-05	6.53	2.00E-05	6.53	6.79E-03	1.48	6.79E-03	1.48	8.16E-03	1.31	8.16E-03	1.31
11113	9.46E-07	4.55	9.45E-07	4.55	2.06E-03	1.77	2.07E-03	1.77	4.04E-03	1.05	4.04E-03	1.05
42097	6.28E-08	4.07	6.25E-08	4.08	2.94E-04	2.92	2.95E-04	2.92	2.00E-03	1.06	2.00E-03	1.06
166966	3.79E-09	4.08	3.73E-09	4.09	6.67E-06	5.50	6.67E-06	5.50	9.09E-04	1.14	9.09E-04	1.14
665302	2.79E-10	3.77	2.67E-10	3.82	2.29E-07	4.88	2.29E-07	4.88	3.20E-04	1.51	3.20E-04	1.51

	$\epsilon = 10^{-2}$				$\epsilon = 10^{-3}$				$\epsilon = 10^{-4}$			
I	IMEX(5,4;1)	rate	IMEX(6,4;1)	rate	(5,4;1)	rate	(6,4;1)	rate	(5,4;1)	rate	(6,4;1)	rate
778	1.48E-03	–	1.47E-03	–	1.80E-02	–	1.80E-02	–	1.93E-02	–	1.93E-02	–
2902	2.00E-05	6.53	2.00E-05	6.53	6.79E-03	1.48	6.79E-03	1.48	8.16E-03	1.31	8.16E-03	1.31
11113	9.45E-07	4.55	9.46E-07	4.55	2.07E-03	1.77	2.06E-03	1.77	4.04E-03	1.05	4.04E-03	1.05
42097	6.26E-08	4.08	6.27E-08	4.08	2.95E-04	2.92	2.94E-04	2.93	2.00E-03	1.06	2.00E-03	1.06
166966	3.73E-09	4.09	3.73E-09	4.09	6.66E-06	5.50	6.65E-06	5.50	9.09E-04	1.14	9.09E-04	1.14
665302	2.65E-10	3.83	2.65E-10	3.83	2.29E-07	4.88	2.29E-07	4.87	3.20E-04	1.51	3.20E-04	1.51

[3] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11** (1973), no. 1, 38–69]. *J. Comput. Phys.*, 135(2):170–186, 1997.

[4] S. Boscarino and L. Pareschi. On the asymptotic properties of IMEX Runge–Kutta schemes for hyperbolic balance laws. *J. Comput. Appl. Math.*, 316:60–73, 2017.

[5] E. Burman and A. Ern. Implicit-explicit Runge-Kutta schemes and finite elements with symmetric stabilization for advection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(4):681–707, 2012.

[6] M. P. Calvo, J. de Frutos, and J. Novo. Linearly implicit Runge-Kutta methods for advection-reaction-diffusion equations. *Appl. Numer. Math.*, 37(4):535–549, 2001.

[7] F. Coquel and P. LeFloch. Convergence of finite difference schemes for conservation laws in several space dimensions: the corrected antidiffusive flux approach. *Math. Comp.*, 57(195): 169–210, 1991.

[8] M. Crouzeix. *Sur l'approximation des équations différentielles linéaires par des méthodes de Runge–Kutta*. PhD thesis, Univ. Paris 6, France, 1975. 192pp.

[9] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numer. Math.*, 35(3):257–276, 1980.

[10] A. Ern and J.-L. Guermond. Invariant-domain-preserving high-order time stepping: I. explicit runge–kutta schemes, 2021. `https://hal.archives-ouvertes.fr/hal-03425367`.

[11] J.-L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Numer. Analysis*, 54(4):2466–2489, 2016.

[12] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas. Second-order invariant domain preserving approximation of the Euler equations using convex limiting. *SIAM J. Sci. Comput.*, 40(5):A3211–A3239, 2018.

[13] J.-L. Guermond, B. Popov, and I. Tomas. Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Comput. Methods Appl. Mech. Engrg.*, 347:143–175, 2019.

[14] J.-L. Guermond, M. Maier, B. Popov, and I. Tomas. Second-order invariant domain preserving approximation of the compressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 375:Paper No. 113608, 17, 2021.

[15] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations. II. Stiff and Differential-algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Second revised edition, paperback.

[16] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 (3):357–393, 1983.

[17] C. A. Kennedy and M. H. Carpenter. Additive Runge-Kutta schemes for convection-diffusion-reaction equations. *Appl. Numer. Math.*, 44(1-2):139–181, 2003.

[18] C. A. Kennedy and M. H. Carpenter. Higher-order additive Runge-Kutta schemes for ordinary differential equations. *Appl. Numer. Math.*, 136:183–205, 2019.

[19] D. Kuzmin and S. Turek. Flux correction tools for finite elements. *Journal of Computational Physics*, 175(2):525–558, 2002.

[20] D. Kuzmin, R. Löhner, and S. Turek. *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Scientific Computation. Springer, 2012. ISBN 9789400740372.

[21] X.-D. Liu and S. Osher. Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes. I. *SIAM J. Numer. Anal.*, 33(2):760–779, 1996.

[22] S. P. Nørsett. Semi explicit Runge–Kutta methods. Technical Report 6/74, Dept. Math. Univ. Trondheim, 1974. 68+7pp.

[23] S. Osher and S. Chakravarthy. High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.*, 21(5):955–984, 1984.

[24] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. In *Recent trends in numerical analysis*, volume 3 of *Adv. Theory Comput. Math.*, pages 269–288. Nova Sci. Publ., Huntington, NY, 2001.

[25] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25(1-2):129–155, 2005.

[26] R. Sanders. A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws. *Math. Comp.*, 51(184):535–558, 1988.

[27] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439 – 471, 1988.

[28] J. M. Varah. Stability restrictions on second order, three level finite difference schemes for parabolic equations. *SIAM J. Numer. Anal.*, 17(2):300–309, 1980.

[29] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31(3):335–362, 1979.

[30] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229(9):3091–3120, 2010.

[31] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 467(2134):2752–2776, 2011.

[32] X. Zhong. Additive semi-implicit Runge-Kutta methods for computing high-speed nonequilibrium reactive flows. *J. Comput. Phys.*, 128(1):19–31, 1996.