SIAM J. SCI. COMPUT.
Vol. 45, No. 5, pp. A2511–A2538

INVARIANT-DOMAIN PRESERVING HIGH-ORDER TIME STEPPING: II. IMEX SCHEMES[*]

ALEXANDRE ERN[†] AND JEAN-LUC GUERMOND[‡]

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 one to optimize the time step restrictions induced by the hyperbolic substep. 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 systems of stiff ordinary differential equations and nonlinear conservation equations.

Key words. time integration, implicit-explicit time integration methods, strong stability preserving methods, conservation equations, hyperbolic systems, high-order method

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

DOI. 10.1137/22M1505025

1. Introduction.

This work is the second part of a project started in [13] whose objective is to develop Runge–Kutta time stepping schemes that are invariant-domain preserving 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 $T > 0$:

$$\partial_t \boldsymbol{u} + \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{g}(\boldsymbol{u}, \nabla \boldsymbol{u}) = \boldsymbol{0} \text{ in } D \times J, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0 \text{ in } D, \tag{1.1}$$

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 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 $\boldsymbol{f}$ and $\mathcal{A} \times \mathbb{R}^{m\times d}$ is the domain of $\boldsymbol{g}$. In the applications we have in mind, these operators have the following structure:

$$\boldsymbol{f}(\boldsymbol{u}) = \nabla\cdot \mathbb{f}(\boldsymbol{u}), \qquad \boldsymbol{g}(\boldsymbol{u}, \nabla \boldsymbol{u}) = \nabla\cdot \mathbb{d}(\boldsymbol{u}, \nabla \boldsymbol{u}) + \boldsymbol{r}(\boldsymbol{u}), \tag{1.2}$$

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section June 23, 2022; accepted for publication (in revised form) June 9, 2023; published electronically September 22, 2023.

https://doi.org/10.1137/22M1505025

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

[†] CERMICS, Ecole des Ponts, 77455, Marne-la-Vallee Cedex 2, France, and INRIA Paris, 75589, Paris, France (alexandre.ern@enpc.fr).

[‡] Department of Mathematics, Texas A&M University, College Station, TX 77843 USA (guermond@math.tamu.edu).

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 the paper to discuss the existence and uniqueness of solutions to (1.1), 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 the paper), then any admissible solution to (1.1) 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 $\mathcal{A}:=[\text{ess inf}_{\boldsymbol{x}\in D}u_0(\boldsymbol{x}),\text{ess sup}_{\boldsymbol{x}\in D}u_0(\boldsymbol{x})]\subset\mathbb{R}$ 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 $\text{ess inf}_{\boldsymbol{x}\in D}s(\boldsymbol{u}_0(\boldsymbol{x}))$. 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 (1.1) is conservation. We assume that there is a matrix $\boldsymbol{C}\in\mathbb{R}^{m'\times m}$ for some $m'\in\{1:m\}$, so that $\boldsymbol{C}\boldsymbol{r}(\boldsymbol{u})=\boldsymbol{0}$. This means that the relaxation process does not affect the variables $\boldsymbol{C}\boldsymbol{u}$. A simple example is when the rows of $\boldsymbol{C}$ are composed of m' line vectors from the Cartesian basis of $\mathbb{R}^m$. Then, again in the absence of perturbations due to the boundary conditions, the following conservation property holds:

$$\int_D \boldsymbol{C}\boldsymbol{u}(t,\cdot)\,\mathrm{d}x=\int_D \boldsymbol{C}\boldsymbol{u}_0\,\mathrm{d}x \qquad \forall t\in J. \tag{1.3}$$

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 (1.1). In this paper, we mainly focus 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, and finite differences). We assume that the ODE system takes the following generic form:

$$\mathbb{M}\partial_t\mathsf{U}=\mathsf{F}(\mathsf{U})+\mathsf{G}(\mathsf{U})\ \ \forall t\in J, \qquad \mathsf{U}(0)=\mathsf{U}_0. \tag{1.4}$$

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 $\mathsf{U}(t)$ takes values in $(\mathbb{R}^m)^I$ where $I\ge 1$ is the number of degrees of freedom (dofs) employed in the space discretization. We set $\mathcal{V}:=\{1:I\}:=\{1,\dots,I\}$ and write $\mathsf{U}(t):=(\mathsf{U}_i(t))_{i\in\mathcal{V}}$. For all $i\in\mathcal{V}$, the local state vector $\mathsf{U}_i(t)=(\mathsf{U}_{p,i}(t))_{p\in\{1:m\}}$ 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 $\mathsf{F}\in C^0(\mathcal{A}^I;(\mathbb{R}^m)^I)$ and $\mathsf{G}\in C^0(\mathcal{A}^I;(\mathbb{R}^m)^I)$ result from the space discretization of the operators $-\boldsymbol{f}$ and $-\boldsymbol{g}$ in (1.1), respectively, and $\mathsf{U}_0\in\mathcal{A}^I$ is an appropriate approximation in space of the initial datum $\boldsymbol{u}_0$. We loosely refer to the mappings F and G as the hyperbolic flux and the parabolic flux, respectively. Assuming that $\mathsf{U}_i(0)\in\mathcal{A}$ for all $i\in\mathcal{V}$, saying that the space approximation is IDP means that

$$\mathbf{U}(t) \in \mathcal{A}^I \qquad \forall t \in J. \tag{1.5}$$

Saying that the method is conservative means that (the mass, m_i, associated with the dof $i \in \mathcal{V}$ is defined in (2.6))

$$\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i(t) = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i(0) \qquad \forall t \in J. \tag{1.6}$$

Using an implicit scheme to discretize in time the ODE system (1.4) 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 [10, 36]. We also refer the reader to [1, 2, 6, 21, 30, 31, 40] 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 [13], we propose in the paper an answer to this question. More precisely, we introduce a technique that makes every IMEX Runge–Kutta (RK) time stepping method IDP and conservative. The resulting schemes are called "IDP-IMEX" schemes. Recall that the technique introduced in [13] makes every explicit RK scheme IDP. The two key ideas of the method consist of externalizing the limiting operation at each stage of the RK scheme and rewriting the RK scheme in incremental form so as to maximize its efficiency. The idea of externalizing the limiter is also considered in [25] for explicit RK schemes and in [32] for diagonally implicit RK schemes, but the central idea of writing the scheme in incremental form and maximizing efficiency is only developed in [13] for explicit RK schemes and in the present work for 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, and we exemplify these notions for the simplest IMEX scheme composed of one forward Euler step followed by one backward Euler step. 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 (3.9) to (3.21), whose properties are stated in Theorem 3.3. In section 4, we review some examples of higher-order IMEX schemes, and we derive four novel examples with optimal efficiency. Finally, in section 5, we present numerical illustrations on systems of stiff ODEs and nonlinear conservation equations.

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; and (iii) exemplify these notions for the Euler 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 (1.4) at the discrete time t^n. The key invariant-domain property we want to achieve is the following:

$$(\mathbf{U}^n \in \mathcal{A}^I) \Longrightarrow (\mathbf{U}^{n+1} \in \mathcal{A}^I) \qquad \forall n \geq 0. \tag{2.1}$$

And we want to achieve (2.1) 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 (the mass m_i is defined in (2.6))

$$\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n}. \tag{2.2}$$

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

$$\mathbf{G}^{\mathrm{lin}}(\mathbf{W};\mathbf{W}) = \mathbf{G}(\mathbf{W}) \quad \forall \mathbf{W} \in \mathcal{A}^I. \tag{2.3}$$

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

$$\mathbb{M}\mathbf{U} - \tau \mathbf{G}^{\mathrm{lin}}(\mathbf{W};\mathbf{U}) = \mathbb{M}\mathbf{V} \tag{2.4}$$

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 section 5 for examples. Owing to the above quasi-linearization process, we reformulate (1.4) 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 $t \in J^n$,

$$\mathbb{M}\partial_t \mathbf{U} = \underbrace{\mathbf{F}(\mathbf{U}) + \mathbf{G}(\mathbf{U}) - \mathbf{G}^{\mathrm{lin}}(\mathbf{U}^n;\mathbf{U})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\mathrm{lin}}(\mathbf{U}^n;\mathbf{U})}_{\text{implicit}}. \tag{2.5}$$

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 $^{\mathrm{L}}$. The second one is high-order accurate and referred to with the superscript $^{\mathrm{H}}$. The low-order scheme is based on a low-order invertible mass matrix $\mathbb{M}^{\mathrm{L}} \in \mathbb{R}^{I\times I}$ and low-order fluxes $\mathbf{F}^{\mathrm{L}}, \mathbf{G}^{\mathrm{L}} : \mathcal{A}^I \to (\mathbb{R}^m)^I$. The high-order scheme is based on a high-order invertible mass matrix $\mathbb{M}^{\mathrm{H}} \in \mathbb{R}^{I\times I}$ and high-order fluxes $\mathbf{F}^{\mathrm{H}}, \mathbf{G}^{\mathrm{H}} : \mathcal{A}^I \to (\mathbb{R}^m)^I$.

We assume that $\mathbb{M}^{\mathrm{H}}$ is symmetric positive-definite with entries $(m_{ij})_{i,j\in\mathcal{V}}$ and $\mathbb{M}^{\mathrm{L}}$ is diagonal with positive 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 $\mathcal{I}(i)$ the stencil at i. The notion of stencil is symmetric, i.e., $j \in \mathcal{I}(i)$ iff $i \in \mathcal{I}(j)$ because $m_{ij} = m_{ji}$. Finally, we assume that

$$m_i := \sum_{j\in\mathcal{V}} m_{ij} = \sum_{j\in\mathcal{V}} m_{ji} \qquad \forall i \in \mathcal{V}, \tag{2.6}$$

and we set $\delta m_{ij} := m_{ij} - m_i \delta_{ij}$. In the finite element terminology, this means that the low-order mass matrix $\mathbb{M}^{\mathrm{L}}$ is the lumped version of the high-order mass matrix $\mathbb{M}^{\mathrm{H}}$. 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 all $i \in \mathcal{V}$.

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

$$\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_i^{\mathrm{H}}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}) \qquad \forall \mathbf{V}\in\mathcal{A}^I, \tag{2.7}$$

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

$$\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{H}}(\mathbf{V}) \qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{2.8}$$

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

$$\mathbf{G}_i^{\mathrm{H}}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}_{ij}^{\mathrm{H}}(\mathbf{V}) + \mathbf{R}_i^{\mathrm{H}}(\mathbf{V}), \qquad \mathbf{D}_{ij}^{\mathrm{H}}(\mathbf{V}) = -\mathbf{D}_{ji}^{\mathrm{H}}(\mathbf{V}) \qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{2.9}$$

For instance, using continuous finite elements with shape functions $\{\varphi_j\}_{j\in\mathcal{V}}$, and assuming for simplicity the parabolic operator to be the Laplacian, we have $\mathbf{D}_{ij}^{\mathrm{H}}(\mathbf{V}) = \int_D \nabla\varphi_j{\cdot}\nabla\varphi_i \mathrm{d}x(\mathbf{V}_j - \mathbf{V}_i) = -\mathbf{D}_{ji}^{\mathrm{H}}(\mathbf{V})$. Consistently with our assumption that $\boldsymbol{Cr}(\boldsymbol{u}) = \mathbf{0}$, we assume that $\boldsymbol{C}\mathbf{R}_i^{\mathrm{H}}(\mathbf{V}) = \mathbf{0}$ for all $i\in\mathcal{V}$.

The quasi-linearization process mentioned in (2.5) 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:

$$\mathbf{G}_i^{\mathrm{L,lin}}(\mathbf{W};\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}_{ij}^{\mathrm{L,lin}}(\mathbf{W};\mathbf{V}) + \mathbf{R}_i^{\mathrm{L,lin}}(\mathbf{W};\mathbf{V}), \tag{2.10}$$

$$\mathbf{G}_i^{\mathrm{H,lin}}(\mathbf{W};\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{D}_{ij}^{\mathrm{H,lin}}(\mathbf{W};\mathbf{V}) + \mathbf{R}_i^{\mathrm{H,lin}}(\mathbf{W};\mathbf{V}), \tag{2.11}$$

$$\mathbf{D}_{ij}^{\mathrm{L,lin}}(\mathbf{W};\mathbf{V}) = -\mathbf{D}_{ji}^{\mathrm{L,lin}}(\mathbf{W};\mathbf{V}), \qquad \mathbf{D}_{ij}^{\mathrm{H,lin}}(\mathbf{W};\mathbf{V}) = -\mathbf{D}_{ji}^{\mathrm{H,lin}}(\mathbf{W};\mathbf{V}), \tag{2.12}$$

$$\boldsymbol{C}\mathbf{R}_i^{\mathrm{L,lin}} = \boldsymbol{C}\mathbf{R}_i^{\mathrm{H,lin}} = \mathbf{0} \qquad \forall i\in\mathcal{V}. \tag{2.13}$$

In conclusion, we consider two versions of the ODE system (2.5) over the time interval $J^n = [t^n, t^{n+1}]$. The first one corresponds to the low-order space discretization:

$$\mathbb{M}^{\mathrm{L}}\partial_t\mathbf{U}^{\mathrm{L}} = \underbrace{\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{\mathrm{L}})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\mathrm{L,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{L}})}_{\text{implicit}}. \tag{2.14}$$

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

$$\mathbb{M}^{\mathrm{H}}\partial_t\mathbf{U}^{\mathrm{H}} = \underbrace{\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{\mathrm{H}}) + \mathbf{G}^{\mathrm{H}}(\mathbf{U}^{\mathrm{H}}) - \mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H}})}_{\text{explicit}} + \underbrace{\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H}})}_{\text{implicit}}. \tag{2.15}$$

Consistently with our assumption on (2.4), we assume that for all $\mathbf{W}\in\mathcal{A}^I$ and all $\mathbf{V}\in(\mathbb{R}^m)^I$, the following problems are well-posed and easy to solve: Find $\mathbf{U}^{\mathrm{L}}, \mathbf{U}^{\mathrm{H}} \in (\mathbb{R}^m)^I$ so that

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L}} - \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W};\mathbf{U}^{\mathrm{L}}) = \mathbb{M}^{\mathrm{L}}\mathbf{V}, \qquad \mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H}} - \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{W};\mathbf{U}^{\mathrm{H}}) = \mathbb{M}^{\mathrm{H}}\mathbf{V}. \tag{2.16}$$

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 schemes. We call this method Euler IMEX. Let us apply it to the low-order ODE system (2.14). Let $\mathbf{U}^n \in \mathcal{A}^I$. The first step is explicit and consists of computing the hyperbolic prediction

$$\mathbf{W}^{\mathrm{L},n} := \big(\mathbb{I} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}\big)(\mathbf{U}^n). \tag{2.17}$$

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

$$\big(\mathbb{I} - \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{\mathrm{L},n};\cdot)\big)(\mathbf{U}^{\mathrm{L},n+1}) = \mathbf{W}^{\mathrm{L},n}. \tag{2.18}$$

Altogether, we have

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} = \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n) + \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{\mathrm{L},n};\mathbf{U}^{\mathrm{L},n+1}). \tag{2.19}$$

This leads us to formulate the following two key structural assumptions.

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

$$\big\{\, \mathbf{V} \in \mathcal{A}^I \,\big\} \Longrightarrow \big\{\, \big(\mathbb{I} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}\big)(\mathbf{V}) \in \mathcal{A}^I \,\big\}, \tag{2.20}$$

$$\big\{\, \mathbf{V} \in \mathcal{A}^I \,\big\} \Longrightarrow \big\{\, \big(\mathbb{I} - \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{G}^{\mathrm{L,lin}}(\mathbf{V};\cdot)\big)^{-1}(\mathbf{V}) \in \mathcal{A}^I \,\big\}. \tag{2.21}$$

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* (2.19) *is well-defined, IDP, and conservative, i.e.,* $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$ *and* $\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^n$.

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 (2.20) to hold. The invariant-domain property of the parabolic step (2.21) 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 (2.15) consists of seeking $\mathbf{U}^{\mathrm{H},n+1} \in (\mathbb{R}^m)^I$ so that

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n) + \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}). \tag{2.22}$$

Similarly to (2.19), the method (2.22) 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, 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 [4, 37, 29, 19]. 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 [33, Thm. 2.1], [8, Thm. 4.3], [26, Thm. 1], and [38, Thm. 2.5]). A well-known limiting method for scalar conservation equations is the so-called flux-corrected transport (FCT) technique of [4] and [37]. The reader is also referred to [23, 24] for extensions. For hyperbolic systems, so-called synchronized algorithms, which still compute limiters based on sequentially determined bounds for each scalar component, are considered in [34] (see also [27] and the references therein for extensions). Nonlinear limiting algorithms enforcing bounds coupling together several components in a nonaffine manner are derived in, e.g., [33, Lem. 3.3], [8, Thm. 4.3], [26, Thm. 2], [39, Lem. 2.4], and [15, 16]. The key idea common to all the above techniques is the decomposition of the flux over the stencils in skew-symmetric components as in (2.7), (2.8), and (2.9).

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 in $\mathbb{R}^{I\times I}$ with the sparsity pattern induced by the stencils $(\mathcal{I}(i))_{i\in\mathcal{V}}$ and coefficients in $[0,1]$. 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 $\mathcal{B} := \ker(\boldsymbol{C}) \subset \mathbb{R}^m$.

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 a 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:*

$$\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij} \in \mathcal{A} \quad \forall i\in\mathcal{V}. \tag{2.23}$$

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

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:*

$$\mathbf{V}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \ell^a_{ij}\mathbf{A}_{ij} + \frac{\tau}{m_i}\ell^b_i\mathbf{B}_i \in \mathcal{A} \quad \forall i\in\mathcal{V}. \tag{2.24}$$

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 by* $\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})$.

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* $(\mathbf{V},\mathbf{A},\mathbf{B}) \in \mathcal{A}^I \times \mathfrak{M} \times \mathcal{B}^I$, *we have*

$$\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\boldsymbol{\ell}^{\rm hyp}(\mathbf{V},\mathbf{A})_i = \sum_{i\in\mathcal{V}} m_i\boldsymbol{C}\mathbf{V}_i, \qquad \sum_{i\in\mathcal{V}} m_i\boldsymbol{C}\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})_i = \sum_{i\in\mathcal{V}} m_i\boldsymbol{C}\mathbf{V}_i. \tag{2.25}$$

Proof. For the parabolic limiter, we have

$$\sum_{i\in\mathcal{V}} m_i\boldsymbol{C}\boldsymbol{\ell}^{\rm par}(\mathbf{V},\mathbf{A},\mathbf{B})_i = \sum_{i\in\mathcal{V}} m_i\boldsymbol{C}\mathbf{V}_i + \tau\boldsymbol{C}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} \ell^a_{ij}\mathbf{A}_{ij},$$

because $\mathbf{B} \in \mathcal{B}^I$ and $\mathcal{B} = \ker(\boldsymbol{C})$. But the symmetry property $\ell^a_{ij} = \ell^a_{ji}$ and the skew-symmetry property $\mathbf{A}_{ij} = -\mathbf{A}_{ji}$ imply that $\sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{I}(i)} \ell^a_{ij}\mathbf{A}_{ij} = \mathbf{0}$, 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:

$$\mathbf{U}^n \underbrace{\xrightarrow{(1)} (\mathbf{W}^{\mathrm{L},n+1}, \mathbf{W}^{\mathrm{H},n+1}) \xrightarrow{(2)}}_{\text{hyperbolic step}} \mathbf{W}^{n+1} \underbrace{\xrightarrow{(3)} (\mathbf{U}^{\mathrm{L},n+1}, \mathbf{U}^{\mathrm{H},n+1}) \xrightarrow{(4)}}_{\text{parabolic step}} \mathbf{U}^{n+1}. \tag{2.26}$$

Let us now give the details of the four steps. We essentially follow Zalesak's limiting strategy [37, eqn. (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

$$\mathbb{M}^{\mathrm{L}}\mathbf{W}^{\mathrm{L},n+1} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n), \tag{2.27}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{\mathrm{H},n+1} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n). \tag{2.28}$$

In step (2), we apply the hyperbolic limiting operator. Subtracting (2.27) from (2.28) and using (2.7) and (2.8), elementary manipulations show that for all $i \in \mathcal{V}$,

$$\mathbf{W}^{\mathrm{H},n+1}_i = \mathbf{W}^{\mathrm{L},n+1}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbf{A}^n_{ij}, \tag{2.29}$$

$$\text{with} \quad \mathbf{A}^n_{ij} := \mathbf{F}^{\mathrm{H}}_{ij}(\mathbf{U}^n) - \mathbf{F}^{\mathrm{L}}_{ij}(\mathbf{U}^n) - \frac{\delta m_{ij}}{\tau}(\mathbf{W}^{\mathrm{H},n+1}_j - \mathbf{U}^n_j - \mathbf{W}^{\mathrm{H},n+1}_i + \mathbf{U}^n_i). \tag{2.30}$$

Notice that $\mathbf{A}^n \in \mathfrak{M}$ is in compliance with Definition 2.4. The conservative IDP hyperbolic high-order update is then obtained by setting

$$\mathbf{W}^{n+1} := \ell^{\mathrm{hyp}}(\mathbf{W}^{\mathrm{L},n+1}, \mathbf{A}^n). \tag{2.31}$$

Parabolic steps (3) and (4). Step (3) consists of computing the low-order and high-order parabolic updates by solving

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} - \tau\mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}) := \mathbb{M}^{\mathrm{L}}\mathbf{W}^{n+1}, \tag{2.32}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} - \tau\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},n+1}) := \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n+1}. \tag{2.33}$$

The quasi-linearization in (2.32) is based on $\mathbf{W}^{n+1}$ to be able to invoke assumption (2.21). The quasi-linearization in (2.33) 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 (2.32) from (2.33) and using (2.10), (2.11), and (2.12), elementary manipulations show that for all $i \in \mathcal{V}$,

$$\mathbf{U}^{\mathrm{H},n+1}_i = \mathbf{U}^{\mathrm{L},n+1}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbf{A}^n_{ij} + \frac{\tau}{m_i}\mathbf{B}^n_i, \tag{2.34}$$

$$\begin{aligned} \text{with} \quad \mathbf{A}^n_{ij} := {} & \mathbf{D}^{\mathrm{H,lin}}_{ij}(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}) - \mathbf{D}^{\mathrm{L,lin}}_{ij}(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}) \\ & - \frac{\delta m_{ij}}{\tau}(\mathbf{U}^{\mathrm{H},n+1}_j - \mathbf{W}^{n+1}_j - \mathbf{U}^{\mathrm{H},n+1}_i + \mathbf{W}^{n+1}_i), \end{aligned} \tag{2.35}$$

$$\text{and} \quad \mathbf{B}^n_i := \mathbf{R}^{\mathrm{H,lin}}_i(\mathbf{U}^n;\mathbf{U}^{\mathrm{H},n+1}) - \mathbf{R}^{\mathrm{L,lin}}_i(\mathbf{W}^{n+1};\mathbf{U}^{\mathrm{L},n+1}). \tag{2.36}$$

Notice that $\mathbf{A}^n \in \mathfrak{M}$ and $\mathbf{B}^n \in \mathcal{B}^I$ in compliance with Definition 2.5. The conservative IDP parabolic high-order update is then obtained by setting

$$\mathbf{U}^{n+1} := \ell^{\mathrm{par}}(\mathbf{U}^{\mathrm{L},n+1}, \mathbf{A}^n, \mathbf{B}^n). \tag{2.37}$$

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* $\boldsymbol{\ell}^{\text{hyp}}$ *and* $\boldsymbol{\ell}^{\text{par}}$ *match Definitions* 2.4 *and* 2.5. *Let* $\mathsf{U}^n \in \mathcal{A}^I$. *Then the above Euler IMEX scheme* (2.27)*–*(2.37) *is well-defined, IDP (i.e., it satisfies* (2.1)*), and conservative (i.e., it satisfies* (2.2)*).*

Proof. (1) Let us first prove the method is well-defined and IDP. Since $\mathsf{U}^n \in \mathcal{A}^I$, invoking (2.20) gives $\mathsf{W}^{\mathrm{L},n+1} \in \mathcal{A}^I$. The definition of the hyperbolic limiter then implies that $\mathsf{W}^{n+1} \in \mathcal{A}^I$. Invoking (2.21) then shows that $\mathsf{U}^{\mathrm{L},n+1}$ is well-defined and $\mathsf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$. Finally, the definition of the parabolic limiter implies that $\mathsf{U}^{n+1} \in \mathcal{A}^I$.

(2) Conservativity follows from the following identities:

$$\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathsf{U}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathsf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathsf{W}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathsf{W}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathsf{U}_i^{n},$$

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 section 2.4 to every 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 the 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:

$$(3.1)\qquad \begin{array}{c|cccccc} 0 & 0 & & & & \\ c_2 & a^{\mathsf{e}}_{2,1} & 0 & & & \\ c_3 & a^{\mathsf{e}}_{3,1} & a^{\mathsf{e}}_{3,2} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ c_s & a^{\mathsf{e}}_{s,1} & a^{\mathsf{e}}_{s,2} & \cdots & a^{\mathsf{e}}_{s,s-1} & 0 \\ \hline & b^{\mathsf{e}}_1 & b^{\mathsf{e}}_2 & \cdots & b^{\mathsf{e}}_{s-1} & b^{\mathsf{e}}_s \end{array} \qquad \begin{array}{c|cccccc} 0 & 0 & & & & \\ c_2 & a^{\mathsf{i}}_{2,1} & a^{\mathsf{i}}_{2,2} & & & \\ c_3 & a^{\mathsf{i}}_{3,1} & a^{\mathsf{i}}_{3,2} & a^{\mathsf{i}}_{3,3} & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ c_s & a^{\mathsf{i}}_{s,1} & a^{\mathsf{i}}_{s,2} & \cdots & a^{\mathsf{i}}_{s,s-1} & a^{\mathsf{i}}_{s,s} \\ \hline & b^{\mathsf{i}}_1 & b^{\mathsf{i}}_2 & \cdots & b^{\mathsf{i}}_{s-1} & b^{\mathsf{i}}_s \end{array}$$

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). Finally, we use the convention that $a^{\mathsf{i}}_{s+1,s+1} := 0$.

We are going to assume that

$$(3.2)\qquad \sum_{l\in\{1:j\}} a^{\mathsf{e}}_{j,l} = \sum_{l\in\{1:j\}} a^{\mathsf{i}}_{j,l} = c_j \qquad \forall j \in \{1 : s\}.$$

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 (3.2) also holds true for $j = s+1$.

3.2. High-order IMEX scheme in incremental form. Following the ideas developed in [13] for ERK schemes, we write the IMEX scheme in incremental form. 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 $c_l - c_k$ such that $c_l - c_k \geq 0$ for all $k \in \{1:l-1\}$:

$$c_l - c_{l'(l)} = \min_{\{k\in\{1:l-1\}\,\mid\, c_l - c_k \geq 0\}} (c_l - c_k) \qquad \forall l \in \{2:s+1\}. \tag{3.3}$$

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 this set 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 optimize the CFL restriction on the time step. For further reference, we define

$$\Delta c^{\max} := \max_{l\in\{2:s+1\}} \big(c_l - c_{l'(l)}\big). \tag{3.4}$$

Notice that $\Delta c^{\max} \geq \frac{1}{s}$ and $\Delta c^{\max} = \frac{1}{s}$ whenever all the stages are equidistributed, i.e., $c_l = \frac{l-1}{s}$, $l \in \{1:s+1\}$. In the rest of the paper, we simply write l' instead of $l'(l)$ to simplify the notation, and we set $\delta c_l := c_l - c_{l'}$, $\delta a^{\mathsf{e}}_{l,k} := a^{\mathsf{e}}_{l,k} - a^{\mathsf{e}}_{l',k}$, $\delta a^{\mathsf{i}}_{l,k} := a^{\mathsf{i}}_{l,k} - a^{\mathsf{i}}_{l',k}$ for all $l \in \{2:s+1\}$ and all $k \in \{1:l-1\}$.

We can now approximate in time the high-order ODE system (2.15) by using the IMEX method defined by the two Butcher tableaux in (3.1). We first set $\mathbf{U}^{n,1} := \mathbf{U}^n$. Then, for all $l \in \{2:s+1\}$, the lth stage of the method consists of computing the following high-order update by using the incremental form of the IMEX scheme:

$$\begin{aligned}\mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} \Big\{ &\delta a^{\mathsf{e}}_{l,k}\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}) + \delta a^{\mathsf{e}}_{l,k}\mathbf{G}^{\mathrm{H}}(\mathbf{U}^{n,k}) \\ &+ (\delta a^{\mathsf{i}}_{l,k} - \delta a^{\mathsf{e}}_{l,k})\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k})\Big\} + \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,l}).\end{aligned} \tag{3.5}$$

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

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} \delta a^{\mathsf{e}}_{l,k}\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}), \tag{3.6}$$

$$\begin{aligned}\mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l} - \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,l}) &:= \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} \\ &+ \tau \sum_{k\in\{1:l-1\}} \Big\{\delta a^{\mathsf{e}}_{l,k}\mathbf{G}^{\mathrm{H}}(\mathbf{U}^{n,k}) + (\delta a^{\mathsf{i}}_{l,k} - \delta a^{\mathsf{e}}_{l,k})\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^n;\mathbf{U}^{n,k})\Big\}.\end{aligned} \tag{3.7}$$

Notice that, consistently with (2.5), the quasi-linearization is done with respect to the initial state $\mathbf{U}^n$ at all stages.

3.3. IDP-IMEX scheme. We now make the scheme (3.6)–(3.7) IDP by proceeding as in section 2.4. Given $\mathbf{U}^n \in \mathcal{A}^I$, we set $\mathbf{U}^{n,1} := \mathbf{U}^n$ and decompose each stage $l \in \{2:s+1\}$ into the following four steps:

$$\mathbf{U}^{n,l'} \underbrace{\xrightarrow{(1)} (\mathbf{W}^{\mathrm{L},l}, \mathbf{W}^{\mathrm{H},l}) \xrightarrow{(2)}}_{\text{hyperbolic step (3.6)}} \mathbf{W}^{n,l} \underbrace{\xrightarrow{(3)} (\mathbf{U}^{\mathrm{L},l}, \mathbf{U}^{\mathrm{H},l}) \xrightarrow{(4)}}_{\text{parabolic step (3.7)}} \mathbf{U}^{n,l}. \tag{3.8}$$

Hyperbolic steps (1) and (2). We proceed as in [13] to make the hyperbolic update (3.6) IDP. We first compute the low-order and high-order hyperbolic updates:

$$\mathbb{M}^{\mathrm{L}}\mathbf{W}^{\mathrm{L},l} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau\mathbf{F}^{\mathrm{L},\mathsf{e}} \quad \text{with} \quad \mathbf{F}^{\mathrm{L},\mathsf{e}} := \delta c_l \mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \tag{3.9}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{W}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau\mathbf{F}^{\mathrm{H},\mathsf{e}} \quad \text{with} \quad \mathbf{F}^{\mathrm{H},\mathsf{e}} := \sum_{k\in\{1:l-1\}} \delta a^{\mathsf{e}}_{l,k}\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{3.10}$$

By proceeding as in (2.29)–(2.30), we write

$$\mathbf{W}^{\mathrm{H},l}_i = \mathbf{W}^{\mathrm{L},l}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbf{A}^{n,l}_{ij} \qquad \forall i\in\mathcal{V}, \tag{3.11}$$

$$\text{with} \quad \mathbf{A}^{n,l}_{ij} := \mathbf{F}^{\mathrm{H},\mathsf{e}}_{ij} - \mathbf{F}^{\mathrm{L},\mathsf{e}}_{ij} - \frac{\delta m_{ij}}{\tau}(\mathbf{W}^{\mathrm{H},l}_j - \mathbf{U}^{n,l'}_j - \mathbf{W}^{\mathrm{H},l}_i + \mathbf{U}^{n,l'}_i). \tag{3.12}$$

Notice that $\mathbf{A}^{n,l}\in\mathfrak{M}$ in compliance with Definition 2.4. Second, using the hyperbolic limiter, we set

$$\mathbf{W}^{n,l} := \ell^{\mathrm{hyp}}(\mathbf{W}^{\mathrm{L},l},\mathbf{A}^{n,l}). \tag{3.13}$$

Parabolic steps (3) and (4). We now compute the low-order and high-order parabolic updates defined by solving the following two problems:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} - \tau\delta c_l \mathbf{G}^{\mathrm{L,lin}}(\mathbf{W}^{n,l};\mathbf{U}^{\mathrm{L},l}) := \mathbb{M}^{\mathrm{L}}\mathbf{W}^{n,l}, \tag{3.14}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} - \tau a^{\mathsf{i}}_{l,l}\mathbf{G}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},l}) := \mathbb{M}^{\mathrm{H}}\mathbf{W}^{n,l} + \tau(\mathbf{D}^{\mathrm{H},\mathsf{i}} + \mathbf{R}^{\mathrm{H},\mathsf{i}}) \tag{3.15}$$

$$\text{with} \quad \mathbf{D}^{\mathrm{H},\mathsf{i}} := \sum_{k\in\{1:l-1\}} \Big\{\delta a^{\mathsf{e}}_{l,k}\mathbf{D}^{\mathrm{H}}(\mathbf{U}^{n,k}) + (\delta a^{\mathsf{i}}_{l,k} - \delta a^{\mathsf{e}}_{l,k})\mathbf{D}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{n,k})\Big\}, \tag{3.16}$$

$$\text{and} \quad \mathbf{R}^{\mathrm{H},\mathsf{i}} := \sum_{k\in\{1:l-1\}} \Big\{\delta a^{\mathsf{e}}_{l,k}\mathbf{R}^{\mathrm{H}}(\mathbf{U}^{n,k}) + (\delta a^{\mathsf{i}}_{l,k} - \delta a^{\mathsf{e}}_{l,k})\mathbf{R}^{\mathrm{H,lin}}(\mathbf{U}^{n};\mathbf{U}^{n,k})\Big\}. \tag{3.17}$$

The update $\mathbf{U}^{n,l}$ is obtained by employing the conservative parabolic limiter and by proceeding as for the Euler IMEX scheme. Subtracting (3.14) from (3.15) yields

$$\mathbf{U}^{\mathrm{H},l}_i = \mathbf{U}^{\mathrm{L},l}_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)} \mathbf{A}^{n,l}_{ij} + \frac{\tau}{m_i}\mathbf{B}^{n,l}_i \qquad \forall i\in\mathcal{V}, \tag{3.18}$$

$$\begin{aligned} \text{with} \quad \mathbf{A}^{n,l}_{ij} := {}& a^{\mathsf{i}}_{l,l}\mathbf{D}^{\mathrm{H,lin}}_{ij}(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},l}) - \delta c_l \mathbf{D}^{\mathrm{L,lin}}_{ij}(\mathbf{W}^{n,l};\mathbf{U}^{\mathrm{L},l}) + \mathbf{D}^{\mathrm{H},\mathsf{i}}_{ij} \\ & - \frac{\delta m_{ij}}{\tau}\big(\mathbf{U}^{\mathrm{H},l}_j - \mathbf{W}^{n,l}_j - \mathbf{U}^{\mathrm{H},l}_i + \mathbf{W}^{n,l}_i\big), \end{aligned} \tag{3.19}$$

$$\text{and} \quad \mathbf{B}^{n,l}_i := a^{\mathsf{i}}_{l,l}\mathbf{R}^{\mathrm{H,lin}}_i(\mathbf{U}^{n};\mathbf{U}^{\mathrm{H},l}) - \delta c_l \mathbf{R}^{\mathrm{L,lin}}_i(\mathbf{W}^{n,l};\mathbf{U}^{\mathrm{L},l}) + \mathbf{R}^{\mathrm{H},\mathsf{i}}_i. \tag{3.20}$$

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

$$\mathbf{U}^{n,l} := \ell^{\mathrm{par}}(\mathbf{U}^{\mathrm{L},l},\mathbf{A}^{n,l},\mathbf{B}^{n,l}). \tag{3.21}$$

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 parabolic update (3.15) 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 (2.21).

Remark 3.2 (complexity). The low-order update (3.14) requires solving a quasi-linear system for all $l \in \{2:s+1\}$. The high-order update (3.15) requires solving a quasi-linear system for all $l \in \{2:s\}$ and amounts to an explicit update for $l = s+1$ because $a^{\mathrm{i}}_{s+1,s+1} = 0$. Thus, the above method requires solving $(2s-1)$ quasi-linear systems over each time interval. The low-order parabolic update has to be computed only if limiting is required. Our experience is that parabolic limiting is infrequently needed if one only enforces global bounds at the end of the parabolic substeps (like positivity of some quantity). Parabolic limiting is nevertheless theoretically required to assert the invariant-domain property as counterexamples can be constructed.

Conclusion. The main motivation for the construction introduced above is the following result. Recall that $\Delta c^{\max}$ is defined in (3.4).

THEOREM 3.3 (s-stage IDP-IMEX). *Let* $\mathbf{U}^n \in \mathcal{A}^I$. *Let Assumption* 2.1 *hold and*

$$\tau \in \Big(0, \frac{\tau^*}{\Delta c^{\max}}\Big]. \tag{3.22}$$

Assume that the limiters $\boldsymbol{\ell}^{\mathrm{hyp}}$ *and* $\boldsymbol{\ell}^{\mathrm{par}}$ *match Definitions* 2.4 *and* 2.5. *Consider the s-stage IMEX scheme composed of the steps* (3.9)–(3.21) *for all* $l \in \{2:s+1\}$. *This scheme is well-defined, IDP (i.e., it satisfies* (2.1)*), and conservative (i.e., it satisfies* (2.2)*.)*

Proof. Assume (3.22) 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 $l = 1$. Now let $l \in \{2:s+1\}$. We make the following observations: The low-order hyperbolic update (3.9) has the same structure as (2.27); the high-order hyperbolic update (3.11)–(3.12) has the same structure as (2.29)–(2.30); the low-order parabolic update (3.14) has the same structure as (2.32); the high-order parabolic update (3.18)–(3.21) has the same structure as (2.34)–(2.37). Hence, we can apply Lemma 2.7 to the scheme (3.9)–(3.21), provided the effective time step $(c_l - c_{l'})\tau$ used in the low-order hyperbolic and parabolic stages (3.9) and (3.14) lies in the interval $(0,\tau^*]$. However, this is the case owing to the assumption (3.22). Then Lemma 2.7 implies that $\mathbf{U}^{n,l}$ is well-defined and is in $\mathcal{A}^I$. This proves the assertion. Moreover, Lemma 2.7 also asserts that

$$\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n,l} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n,l'}.$$

An induction argument gives $\sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n,l} = \sum_{i\in\mathcal{V}} m_i \boldsymbol{C}\mathbf{U}_i^{n}$ for all $l \in \{1:s+1\}$, and the conservation property (2.2) follows from $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1}$. ∎

Following [35] (see also [13, Def. 2.2]), the quantity

$$c_{\mathrm{eff}} := \frac{1}{s\Delta c^{\max}} \tag{3.23}$$

is called the *efficiency ratio* of the s-stage IMEX scheme. By construction, we have $c_{\mathrm{eff}} \le 1$. Theorem 3.3 shows that the IMEX scheme is IDP for all $\tau \in (0, c_{\mathrm{eff}} s\tau^*]$. Hence, computational efficiency increases with the efficiency ratio. In particular, the largest time step allowed is $s \times \tau^*$ when $c_{\mathrm{eff}} = 1$. The optimal value $c_{\mathrm{eff}} = 1$ is attained when the coefficients c_j are equidistributed, 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 four 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 $s \geq p$ stages and are described by the Butcher tableaux introduced in (3.1). In what follows, we use the terminology IMEX$(s,p;c_{\text{eff}})$ for IMEX schemes 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; see (4.18).
(2) IMEX$(4,3;1)$, singly diagonal, L-stable implicit part; see (4.19).
(3) IMEX$(5,4;1)$, singly diagonal, L-stable implicit part; see (4.20).
(4) IMEX$(6,4;1)$, singly diagonal, L-stable implicit part; see (4.21).

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, $p \geq 0$, 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. The identity matrix of order s is denoted by I_s.

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

$$BA^{r-1}C^{q-1} = \frac{(q-1)!}{(q-1+r)!} \qquad \forall r \in \{1:p\},\ \forall q \in \{1:p-r+1\}, \tag{4.1}$$

where A stands either for A^{e} or for A^{i}. 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 $p \geq 2$. Moreover, using $b_i^{\mathsf{e}} = b_i^{\mathsf{i}} =: b_i$ and $c_i^{\mathsf{e}} = c_i^{\mathsf{i}} =: c_i$ for all $i \in \{1:s\}$ ensures that the IMEX coupling conditions for second- and third-order schemes are satisfied; see [31, eqns. (8) and (10)]. More coupling conditions must be added for IMEX schemes to be of order $p \geq 4$; see [22, Tab. 4].

The consistency properties of IMEX methods are reviewed in [31, sect. 2.1] and [22, sect. 2.2]. The analysis therein is based on the following simplifying assumption (see (3.2)), which we systematically enforce:

$$A^{\mathsf{e}}U = C, \qquad A^{\mathsf{i}}U = C. \tag{4.2}$$

In addition to (4.2), the (linear order) conditions needed to achieve second order are

$$BU = 1, \qquad BC = \tfrac{1}{2}. \tag{4.3}$$

The other conditions, $BA^{\mathsf{e}}U = BA^{\mathsf{i}}U = \frac{1}{2}$, are satisfied owing to (4.2) and (4.3).

In addition to (4.2)–(4.3), the conditions needed to achieve third-order accuracy are

$$BC^2 = \tfrac{1}{3}, \qquad BA^{\mathsf{e}}C = BA^{\mathsf{i}}C = \tfrac{1}{6}. \tag{4.4}$$

The other conditions, $B(A^{\mathsf{e}})^2U = B(A^{\mathsf{i}})^2U = \frac{1}{6}$, are satisfied owing to (4.2) and (4.4).

In addition to (4.2), (4.3), (4.4), the conditions needed to achieve fourth-order accuracy are the linear order conditions

$$BC^3 = \tfrac{1}{4}, \qquad BA^{\mathsf{e}}C^2 = BA^{\mathsf{i}}C^2 = \tfrac{1}{12}, \qquad B(A^{\mathsf{e}})^2C = B(A^{\mathsf{i}})^2C = \tfrac{1}{24} \tag{4.5}$$

plus the nonlinear order condition

$$(B\odot C)A^{\mathsf{e}}C = (B\odot C)A^{\mathsf{i}}C = \tfrac{1}{8} \tag{4.6}$$

and the coupling condition

$$BA^{\mathsf{e}}A^{\mathsf{i}}C = BA^{\mathsf{i}}A^{\mathsf{e}}C = \tfrac{1}{24}. \tag{4.7}$$

The other conditions, $B(A^{\mathsf{e}})^3U = B(A^{\mathsf{i}})^3U = \frac{1}{24}$, follow from (4.2) and (4.5).

Finally, in addition to (4.2)–(4.7), we are also going to make use of the fifth-order linear order conditions for a six-stage, fourth-order method,

$$\begin{aligned} &BC^4 = \tfrac{1}{5}, && BA^{\mathsf{e}}C^3 = BA^{\mathsf{i}}C^3 = \tfrac{1}{20},\\ &B(A^{\mathsf{e}})^2C^2 = B(A^{\mathsf{i}})^2C^2 = \tfrac{1}{60}, && B(A^{\mathsf{e}})^3C = B(A^{\mathsf{i}})^3C = \tfrac{1}{120}. \end{aligned} \tag{4.8}$$

The other linear order conditions, $B(A^{\mathsf{e}})^4U = B(A^{\mathsf{i}})^4U = \frac{1}{120}$, follow from (4.2) and (4.8).

Stability. The amplification function associated with a DIRK scheme is

$$R(z) := 1 + zB(I_s - zA^{\mathsf{i}})^{-1}U, \qquad z \in \mathbb{C}. \tag{4.9}$$

Recall that an RK scheme is said to be A-stable if $|R(z)| \le 1$ for all $z \in \mathbb{C}$ s.t. $\Re(z) \le 0$ (see [18, Def. IV.3.3]). A scheme is said to be L-stable if it is A-stable and $R(t) \to 0$ as $t \to -\infty$ (see [18, 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

$$A^{\mathsf{i}} = \begin{pmatrix} 0 & 0 \\ \alpha & \tilde{A} \end{pmatrix}, \qquad B = (\beta, \tilde{B}), \tag{4.10}$$

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 (4.9) can be rewritten as

$$R(z) = 1 + z\beta + z\tilde{B}(I_{s-1} - z\tilde{A})^{-1}(\tilde{U} + z\alpha), \tag{4.11}$$

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:

$$\beta = \tilde{B}\tilde{A}^{-1}\alpha, \qquad \tilde{B}\tilde{A}^{-1}\tilde{U} + \tilde{B}\tilde{A}^{-2}\alpha = 1. \tag{4.12}$$

Notice that the first condition in (4.12) 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 corresponding Butcher tableaux are

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \end{array} \tag{4.13}$$

We have $l'(l) = l-1$ for all $l \in \{2:3\}$, and $c_{\text{eff}} = \frac{1}{2}$. 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

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 1 & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline 1 & 0 & 1 \end{array} \tag{4.14}$$

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 [2, sect. 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

$$\begin{array}{c|ccc} 0 & 0 & & \\ \gamma & \gamma & 0 & \\ 1 & \delta & 1-\delta & 0 \\ \hline 1 & 0 & 1-\gamma & \gamma \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1 & 0 & 1-\gamma & \gamma \\ \hline 1 & 0 & 1-\gamma & \gamma \end{array} \tag{4.15}$$

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 $l'(l) = l-1$ for all $l \in \{2:6\}$ and $\Delta c^{\max} = \frac{1}{4}$. 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 [17] to solve the compressible Navier–Stokes equations (the method SSPRK$(3,3)$ 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 [28, 9] 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 [2, sect. 2.4])

$$\begin{array}{c|ccc} 0 & 0 & & \\ \gamma & \gamma & 0 & \\ 1-\gamma & \gamma-1 & 2-2\gamma & 0 \\ \hline 1 & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1-\gamma & 0 & 1-2\gamma & \gamma \\ \hline 1 & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \tag{4.16}$$

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

$$R(z) = \frac{1 + (1-2\gamma)z + (\frac{1}{3} - \gamma)z^2}{(1-\gamma z)^2}. \tag{4.17}$$

The zero-padded ESDIRK scheme is A-stable, but not L-stable because we only have $\lim_{t\to-\infty} R(t) = 1 - \sqrt{3} \approx -0.73205$. 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. We first request that the EDIRK scheme have 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 $a^{\mathsf{i}}_{1,1} = 0$, this leaves five entries to be determined. Four equations can be enforced: two from Butcher's simplifying assumption (4.2) involving A^{i} (there are two equations corresponding to the rows $i \in \{2,3\}$ since the row corresponding to $i = 1$ is trivial), one is the (linear order) condition $BA^{\mathsf{i}}C = \frac{1}{6}$ stated in (4.4) (the remaining linear order conditions are already satisfied), and one is the first stability condition in (4.12). One can show that it is not possible to enforce the second equality in (4.12) (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 explicitly enforced. The Butcher arrays of the ERK and ESDIRK schemes are

$$\begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3} & 0 & \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3}-\gamma & \gamma & \\ \frac{2}{3} & \gamma & \frac{2}{3}-2\gamma & \gamma \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \tag{4.18}$$

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 (4.17). Hence, the amplification functions of the methods described by the implicit Butcher tableaux in (4.16) and (4.18) are identical, but the efficiency of (4.16) is only $c_{\text{eff}} = \frac{1}{3}\gamma \approx 0.26$, whereas that of the new method (4.18) is $c_{\text{eff}} = 1$.

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: nine entries in the matrix A^{i} and the four components of the vector b. We enforce Butcher's simplifying assumption (4.2) (three equations), the (linear) order conditions (4.3) and (4.4) (four equations), and the two conditions in (4.12) to achieve L-stability. This gives nine equations. We additionally require that the scheme be singly diagonal, giving the two additional equations $a^{\mathsf{i}}_{2,2} = a^{\mathsf{i}}_{3,3} = a^{\mathsf{i}}_{4,4}$, and that $BC^3 = \frac{1}{4}$ (this is the first of the fourth-order (linear) conditions in (4.5)). The resulting underdetermined set of nonlinear equations (12 equations, 13 unknowns) is solved using `julia` with 10^{-15} tolerance. The following L-stable ESDIRK scheme is found:

(4.19a)

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

The companion ERK method that shares the same set of coefficients c_j and b_j has already been proposed in [13]. The six coefficients of the matrix A^{e} are obtained by enforcing the fourth-order linear consistency conditions (three equations from (4.2), one from (4.4), and two from (4.5)). 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:

$$\begin{array}{c|cccc}
0 & 0 & & & \\
\frac{1}{4} & \frac{1}{4} & 0 & & \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \\
\frac{3}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 \\
\hline
1 & 0 & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3}
\end{array} \tag{4.19b}$$

Near the origin along the imaginary axis, we have $|R^{\mathsf{e}}(i\epsilon)| = 1 + \rho^{\mathsf{e}}_6 \epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho^{\mathsf{e}}_6 = -2B(A^{\mathsf{e}})^4 C + 2B(A^{\mathsf{e}})^3 C - B(A^{\mathsf{e}})^2 C + \frac{1}{36} = -\frac{1}{72}$.

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 [2, sect. 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 [7]. A method combining a five-stage ERK method with a five-stage L-stable DIRK method is described in [2, sect. 2.8] (the DIRK method actually has 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 [22, sects. 3.2 and 3.3], but the efficiency ratio of these methods is not 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 (4.2) (four equations), the (linear) order conditions (4.3), (4.4), and (4.5) (seven equations), the (nonlinear) order condition (4.6) (one equation), and the two conditions in (4.12) to achieve 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 underdetermined set of nonlinear equations (18 equations, 19 unknowns) is solved using `julia` with 2×10^{-16} tolerance. The following solution is found:

(4.20a)

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

$$
\begin{array}{c|ll}
\frac{4}{5} & \dots & 0.57281606248213512 \\
\hline
1 & \dots & 0.35321654878641495
\end{array}
$$

This ESDIRK scheme is L-stable. Along the imaginary axis near the origin, we have $|R^{\mathsf{i}}(\mathsf{i}\epsilon)| = 1 + \rho^{\mathsf{i}}_6 \epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho^{\mathsf{i}}_6 \approx -0.0846$, where $\rho^{\mathsf{i}}_6 = -2B(A^{\mathsf{i}})^4 C + 2B(A^{\mathsf{i}})^3 C - \frac{1}{72}$.

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 (4.2) (four equations) and the (linear) order conditions (4.4) and (4.5) 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 (4.6) (one equation), and the two coupling conditions (4.7). 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:

(4.20b)

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

$$
\begin{array}{c|ll}
\frac{4}{5} & \dots & 0 \\
\hline
1 & \dots & 0.35321654878641495
\end{array}
$$

We have $|R^{\mathsf{e}}(\mathsf{i}\epsilon)| = 1 + \rho^{\mathsf{e}}_6 \epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho^{\mathsf{e}}_6 \approx -0.0148$ along the imaginary axis near the origin, where $\rho^{\mathsf{e}}_6 = 2B(A^{\mathsf{e}})^3 C - \frac{1}{72}$ (notice that $2B(A^{\mathsf{e}})^4 C = 0$).

4.4.2. Six-stage schemes. A six-stage, fourth-order method with L-stable (actually, stiffly accurate) implicit scheme is designed in [7] using an L-stable, six-stage (actually five distinct stages) fourth-order SDIRK method from [18]. The explicit tableau is given by equation (14) in [7], and the implicit tableau is given by equation (IV.6.16) in [18] (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 (4.2) (five equations), the (linear) order conditions (4.3), (4.4), (4.5) (seven equations), the (nonlinear) order condition (4.6) (one equation), and the two conditions in (4.12) to achieve L-stability. We also enforce the fifth-order linear order conditions (4.8) (four 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 $b_4 = 0.47$ (see below), giving five 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:

(4.21a)

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

$$\begin{array}{c|ccc}
\frac{4}{6} & \dots & 0.2780538411364528 & \\
\frac{5}{6} & \dots & -0.3927913893208868 & 0.2780538411364528 \\
\hline
1 & \dots & -0.285 & 0.467
\end{array}$$

We have $|R^{\mathsf{i}}(\mathsf{i}\epsilon)| = 1 + \rho^{\mathsf{i}}_6 \epsilon^6 + \mathcal{O}(\epsilon^8)$ along the imaginary axis near the origin, with $\rho^{\mathsf{i}}_6 \approx -1.06 \times 10^{-3}$ (recall that $\rho^{\mathsf{i}}_6 := -2B(A^{\mathsf{i}})^4 C + 2B(A^{\mathsf{i}})^3 C - \frac{1}{72} = -2B(A^{\mathsf{i}})^4 C + \frac{1}{360}$).

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 (4.2) (five equations), the (linear) order conditions (4.4), (4.5) involving the matrix A^{e} (three equations), the (nonlinear) order condition (4.6) (one equation), and the two coupling conditions (4.7). This gives 11 equations. We also enforce the fifth-order linear order conditions (4.8) (three equations). In total, we have 14 equations. The resulting underdetermined set of nonlinear equations (14 equations, 15 unknowns) is solved using `julia`. The following solution is found with 4×10^{-16} tolerance:

(4.21b)

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

We have $|R^{\mathsf{e}}(\mathsf{i}\epsilon)| = 1 + \rho^{\mathsf{e}}_6 \epsilon^6 + \mathcal{O}(\epsilon^8)$ along the imaginary axis near the origin, with $\rho^{\mathsf{e}}_6 \approx -9.67 \times 10^{-5}$, where $\rho^{\mathsf{e}}_6 := -2B(A^{\mathsf{e}})^4 C + 2B(A^{\mathsf{e}})^3 C - \frac{1}{72} = -2B(A^{\mathsf{e}})^4 C + \frac{1}{360}$. The reason for having set $b_4 = 0.47$ is to make ρ^{e}_6 negative.

5. Numerical illustrations. We illustrate the IMEX methods proposed in the paper on systems of stiff ODEs and nonlinear conservation equations. We start with convergence tests on an ODE system. Then we solve a scalar nonlinear conservation equation with hyperbolic and parabolic fluxes. We finish with one- and two-dimensional simulations of the compressible Navier–Stokes equations.

5.1. Convergence tests. We test the convergence properties of the new IMEX methods proposed in the paper. Following [22, sect. 5.1], we consider the 2×2 ODE system

$$\partial_t y_1(t) = -2y_1 + \epsilon^{-1}(y_2^2 - y_1), \qquad \partial_t y_2(t) = y_1 - y_2 - y_2^2, \tag{5.1}$$

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 $\mathsf{U} := (\mathsf{u}_1, \mathsf{u}_2)^\mathsf{T}$ the approximate solution produced by the IMEX methods. Referring the reader to (2.5) for the notation, we set $\mathbb{M} := \mathbb{I}_2$, where $\mathbb{I}_2$ is the 2×2 identity matrix, and

$$\mathsf{F}(\mathsf{U}) := (-2\mathsf{u}_1, \mathsf{u}_1 - \mathsf{u}_2 - \mathsf{u}_2^2)^\mathsf{T}, \quad \mathsf{G}(\mathsf{U}) := (\epsilon^{-1}(\mathsf{u}_2^2 - \mathsf{u}_1), 0)^\mathsf{T}, \quad \mathsf{G}^{\text{lin}}(\mathsf{W}, \mathsf{U}) := \mathsf{G}(\mathsf{U}). \tag{5.2}$$

Notice that G^{lin} is not linear, but solving the problem (2.4) is simple:

$$(\mathbb{I} - \tau \mathsf{G}^{\text{lin}}(\mathsf{W}, \cdot))^{-1}(\mathsf{W}) = (\tfrac{1}{\epsilon+\tau}(\epsilon \mathsf{w}_1 + \tau \mathsf{w}_2^2), \mathsf{w}_2)^\mathsf{T}. \tag{5.3}$$

We test the methods 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 second-order method IMEX$(2, 2; 1)$ delivers second-order accuracy uniformly with respect to ϵ for both y_1 and y_2 (not shown here for brevity). The errors versus the time step τ are reported in Figure 5.1 for the methods IMEX$(3, 3; 1)$, IMEX$(4, 3; 1)$, IMEX$(5; 4; 1)$, and IMEX$(6, 4; 1)$. 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. We observe the optimal order for the four 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 preasymptotic range (i.e., $\epsilon < \tau$) is well documented in the literature, and we refer the reader to, e.g., [5] for an analysis.

5.2. Nonlinear scalar conservation equation. In this section, we illustrate the method on the scalar nonlinear conservation equation

$$\partial_t u + \nabla \cdot \boldsymbol{f}(u) - \epsilon \Delta u = 0, \quad \boldsymbol{x} \in D_\infty, \ t > 0, \tag{5.4}$$

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

$$u_0(\boldsymbol{x}) := \mu + \delta \tanh\left(\tfrac{\delta}{\epsilon}(x - x_0)\right), \quad \mu := \tfrac{1}{2}(u_L + u_R), \quad \delta := \tfrac{1}{2}(u_R - u_L). \tag{5.5}$$

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

$$v(\boldsymbol{x}, t) = u_0(\boldsymbol{x} - \boldsymbol{s}t) \quad \text{with} \quad \boldsymbol{s} := (s, 0). \tag{5.6}$$

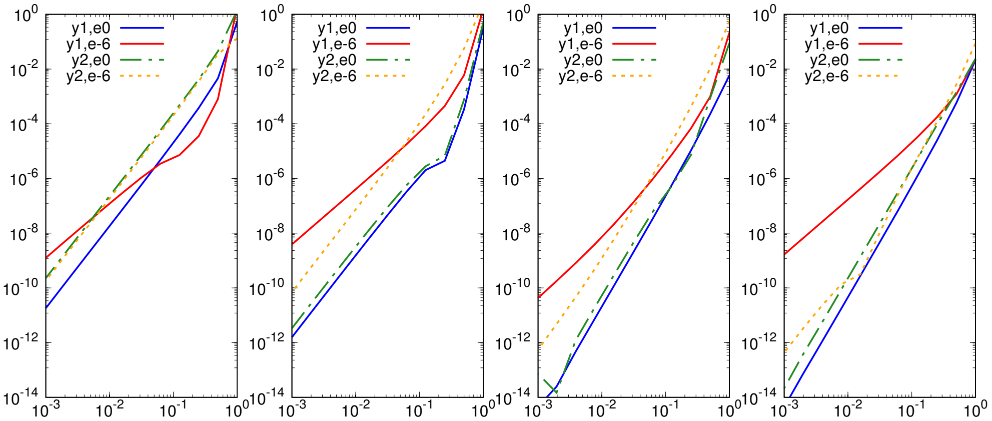

Fig. 5.1. *Convergence test on problem* (5.1). *Left to right: third-order methods* IMEX(3, 3; 1) *and* IMEX(4, 3; 1) *(with the Butcher tableaux* (4.18) *and* (4.19)*); fourth-order methods* IMEX(5; 4; 1) *and* IMEX(6, 4; 1) *(with the Butcher tableaux* (4.20) *and* (4.21)*).*

The method described in the 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 [14]. The high-order method and the limiting are described in [15, 16]. 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 [15, sect. 4.7.1] and [16, sect. 7.6].

We set $u_L := -1$ and $u_R := 1$, so that the solution to (5.4) 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 $\partial D_{\rm D}^{\rm hyp} = \partial D_{\rm D}^{\rm par} := \{x_L, x_R\} \times (y_B, y_T)$ and $\partial D_{\rm N}^{\rm par} := (x_L, x_R) \times \{y_B, y_T\}$ (it happens here that $\partial D_{\rm D}^{\rm hyp} = \partial D_{\rm D}^{\rm par}$ because $1 - 2u_L > 0$ and $1 - 2u_R < 0$). At each stage of the IMEX method, Dirichlet boundary conditions are enforced at $\partial D_{\rm D}^{\rm hyp}$ for the hyperbolic subproblems and at $\partial D_{\rm D}^{\rm par}$ for the parabolic subproblems. Homogeneous Neumann conditions are weakly enforced on $\partial D_{\rm N}^{\rm par}$ for the parabolic subproblems. The enforcement of the Dirichlet 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

$$\tau := \mathrm{CFL} \times s \times \tau^*, \tag{5.7}$$

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 (2.20). 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 $\epsilon = 2 \times 10^{-n}$, $n \in \{2, 3, 4\}$. 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.

TABLE 5.1

Problem (5.4) *for* $\epsilon = 2 \times 10^{-n}$, $n \in \{2,3,4\}$. $\mathbb{P}_1$ *finite elements. Relative 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(2,2;1)	rate	(2,2;1)	rate	(2,2;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

We show the errors and the convergence rates for continuous $\mathbb{P}_1$ elements in Table 5.1. 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 superconvergence 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.

5.3. Compressible Navier–Stokes equations. We consider the compressible Navier–Stokes equations in the domain $D \subset \mathbb{R}^d$, where d is the space dimension. The Cartesian basis of $\mathbb{R}^d$ is denoted by $\{\boldsymbol{e}_k\}_{k\in\{1:d\}}$. The system is equipped with the γ-law. Let μ be the shear viscosity, λ the bulk viscosity, κ the thermal conductivity, $c_V = \frac{1}{\gamma-1}$ the heat capacity at constant volume, $c_P = \gamma c_V$ the heat capacity at constant pressure, and $P_r := \frac{\mu c_P}{\kappa}$ the Prandtl number. Denoting by $\boldsymbol{v}$ the velocity and $\mathbb{e}(\boldsymbol{v})$ the strain tensor, the Newtonian viscous stress tensor is $\mathbb{s}(\boldsymbol{v}) := 2\mu\mathbb{e}(\boldsymbol{v}) - \lambda(\nabla\cdot\boldsymbol{v})\mathbb{I}_{d\times d}$. Denoting by e the specific internal energy and using Fourier's law, the heat flux is $\boldsymbol{q}(e) := -\frac{\kappa}{c_V}\nabla e$.

The approximation in space of the explicit part of the problem, i.e., the Euler equations, is done exactly as in [17]. The limiting operators and the bounds used for the limiting are defined in [15, sects. 4.4 and 4.6]. The relaxation of the bounds guaranteeing high-order convergence is done as explained in [15, sect. 4.7.1] and [16, sect. 7.6]. The fundamental difference with [15] is that the hyperbolic limiting is now done as explained in (3.9)–(3.13). We use Lagrange finite elements for the

Table 5.2

Problem (5.4) *for* $\epsilon = 2\times 10^{-n}$, $n \in \{2,3,4\}$. $\mathbb{P}_3$ *finite elements. Relative 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 (2,2;1)	rate	(2,2;1)	rate	(2,2;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

approximation in space in the tests reported below (tests are done with $\mathbb{P}_1$ and $\mathbb{P}_3$ simplicial elements).

As there are some differences with [17, sect. 5] regarding the approximation of the parabolic part of the problem, we now give some details on the implementation of the parabolic substeps. Let $\{\varphi_i\}_{i\in\mathcal{V}}$ be the scalar Lagrange shape functions that are used for the space approximation. Let $\mathbb{M}$ be the mass matrix with entries $\int_D \varphi_i\varphi_j \mathrm{d}x$. Let us denote by ϱ_i, $\boldsymbol{\mathsf{M}}_i$, $\boldsymbol{\mathsf{V}}_i := \frac{1}{\varrho_i}\boldsymbol{\mathsf{M}}_i$, E_i, and $\mathsf{e}_i := \frac{1}{\varrho_i}\mathsf{E}_i - \frac{1}{2}\|\boldsymbol{\mathsf{V}}_i\|_{\ell^2}^2$ the density, momentum, velocity, total energy, and specific internal energy degrees of freedom at $i\in\mathcal{V}$. Let $l\in\{2:s+1\}$ be the index of the RK stage. Let $\varrho_i^{\mathsf{W},l}$, $\boldsymbol{\mathsf{M}}_i^{\mathsf{W},l}$, $\boldsymbol{\mathsf{V}}_i^{\mathsf{W},l}$, $\mathsf{E}_i^{\mathsf{W},l}$, and $\mathsf{e}_i^{\mathsf{W},l}$ be the density, momentum, velocity, total energy, and specific internal associated with the explicit hyperbolic update $\boldsymbol{\mathsf{W}}^{n,l}$ (see (3.13)) at the lth stage and at the dof $i\in\mathcal{V}$. Let us now describe how the high-order parabolic update $\boldsymbol{\mathsf{U}}^{\mathrm{H},l}$ (see (3.15)) is computed. We use a superscript ^{l} to denote the density, momentum, velocity, total energy, and specific energy associated with $\boldsymbol{\mathsf{U}}^{\mathrm{H},l}$. The update of the density is done by setting $\varrho_i^l = \varrho_i^{\mathsf{W},l}$. The high-order update of the kth Cartesian component of the momentum ($k\in\{1:d\}$) and of the total energy is done as follows:

$$(\mathbb{M}\boldsymbol{\mathsf{M}}^l)_{k,i} + \tau a_{ll}^{\mathsf{i}} a(\boldsymbol{v}_h^l, \varphi_i\boldsymbol{e}_k) = (\mathbb{M}\boldsymbol{\mathsf{M}}^{\mathsf{W},l})_{k,i} - \sum_{r\in\{1:l-1\}} \tau a_{l,r}^{\mathsf{i}} a(\boldsymbol{v}_h^r, \varphi_i\boldsymbol{e}_k), \tag{5.8a}$$

$$\begin{aligned}(\mathbb{M}\mathsf{E}^l)_i + \tau a_{ll}^{\mathsf{i}} b(e_h^l, \varphi_i) &= (\mathbb{M}\mathsf{E}^{\mathsf{W},l})_i \\ &\quad - \sum_{r\in\{1:l-1\}} \tau a_{l,r}^{\mathsf{i}} b(e_h^r, \varphi_i) - \sum_{r\in\{1:l\}} \tau a_{l,r}^{\mathsf{i}} c(\boldsymbol{v}_h^r, \varphi_i)\mathrm{d}x,\end{aligned} \tag{5.8b}$$

where $\boldsymbol{v}_h^r := \sum_{i\in\mathcal{V}} \boldsymbol{\mathsf{V}}_i^r\varphi_i$, $e_h^r := \sum_{i\in\mathcal{V}} \mathsf{e}_i^r\varphi_i$, $a(\boldsymbol{v},\boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}){:}\nabla\boldsymbol{w}\mathrm{d}x$, $b(e,q) := \int_D \frac{\kappa}{c_v}\nabla e{\cdot}\nabla q\mathrm{d}x$, and $c(\boldsymbol{v},q) := \int_D \mathbb{s}(\boldsymbol{v}){:}(\boldsymbol{v}\otimes\nabla q)\mathrm{d}x$. Let $\mathbb{M}^{\varrho,l}$ be the matrix with entries $m_{ij}\varrho_j^{\mathsf{W},l}$, so that $(\mathbb{M}\boldsymbol{\mathsf{M}}^l)_{k,i} = (\mathbb{M}^{\varrho,l}\boldsymbol{\mathsf{V}}^l)_{k,i}$. This identity implies that (5.8a) is an update for the velocity. Similarly we have $(\mathbb{M}\mathsf{E}^l)_i = \mathbb{M}^{\varrho,l}(\mathsf{e}^l + \frac{1}{2}\|\boldsymbol{\mathsf{V}}^l\|_{\ell^2}^2)$, which implies that (5.8b) is an update for the internal energy. Notice that (5.8a)–(5.8b) is nonlinear, but solving this problem only requires linear solves: one first obtains the velocity by solving (5.8a), then one updates the internal energy by solving (5.8b) where the nonlinear terms depending on $\boldsymbol{v}_h^l$ are already known. The momentum and the total energy are updated by setting $\mathbf{M}_i^l = \varrho_i^l \mathbf{V}_i^l$ and $\mathsf{E}_i^l = \varrho_i^l(\mathsf{e}_i^l + \frac{1}{2}\|\mathbf{V}_i^l\|_{\ell^2}^2)$.

If limiting is needed, the low-order parabolic update $\mathbf{U}^{\mathrm{L},l}$ (see (3.14)) is computed as follows: Using again the superscript ^{l} to denote the density, momentum, velocity, total energy, and specific energy associated with $\mathbf{U}^{\mathrm{L},l}$, we set $\varrho_i^l = \varrho_i^{\mathsf{W},l}$ and then

$$
\begin{aligned}
&\text{(5.9a)} \quad m_i \varrho_i^{\mathsf{W},l} \mathbf{V}_{k,i}^l + \tau\delta c_l\, a(\boldsymbol{v}_h^l, \varphi_i \boldsymbol{e}_k) = m_i \mathbf{M}_{k,i}^{\mathsf{W},l},\\
&\text{(5.9b)} \quad m_i \varrho_i^{\mathsf{W},l} \mathsf{e}_i^l + \tau\delta c_l b(e_h^l, \varphi_i) = m_i(\varrho \mathsf{e})_i^{\mathsf{W},l} + \tfrac{m_i \varrho_i^{\mathsf{W},l}}{2}\|\mathbf{V}_i^l - \mathbf{V}_i^{\mathsf{W},l}\|_{\ell^2}^2 + \tau\delta c_l d(\boldsymbol{v}_h^l, \varphi_i),
\end{aligned}
$$

with $d(\boldsymbol{v}, q) := \int_D \mathbb{s}(\boldsymbol{v}) : \mathbb{e}(\boldsymbol{v}) q \mathrm{d}x$. One verifies as in [17, Thm. 5.5] that this update is conservative and IDP under the acute angle condition on the mesh. If necessary, the limiting on the internal energy is done as is [17, sect. 5.3].

5.3.1. One-dimensional simulations. We test the accuracy of the method described above by reproducing the one-dimensional exact solution proposed in [3]. (A partial English translation of [3] and other exact solutions are found in [20].) This test is fully described in [17, sect. 7.2]. We use $\gamma = 1.4$, $\mu = 10^{-2}$, $\lambda = 0$, $P_r := \frac{\mu c_P}{\kappa} = \frac{3}{4}$ (this gives $\kappa = \frac{14}{3} \times 10^{-2}$). The computational domain is $D := [-0.5, 1]$. The final time is $T = 3$. The distance traveled by the viscous shock is 0.6. For every CFL number, the time step is computed by means of the definition (5.7). We compute a consolidated error indicator at the final time by adding the relative error in the L^1-norm on the density, momentum, and total energy as follows:

$$
\delta_1(T) := \frac{\|(\rho_h - \rho)(T)\|_{L^1}}{\|\rho(T)\|_{L^1}} + \frac{\|(\boldsymbol{m}_h - \boldsymbol{m})(T)\|_{\boldsymbol{L}^1}}{\|\boldsymbol{m}(T)\|_{\boldsymbol{L}^1}} + \frac{\|(E_h - E)(T)\|_{L^1}}{\|E(T)\|_{L^1}}. \tag{5.10}
$$

We show in Figure 5.2 the consolidated error as a function of the CFL number for the methods IMEX$(2, 2; \frac{1}{2})$ (blue, - - -), IMEX$(2, 2; 1)$ (blue, ——), IMEX$(3, 3; 0.26)$ (red, - - -), IMEX$(3, 3; 1)$ (red, ——), IMEX$(4, 3; 1)$ (red, ······), IMEX$(5, 4; 1)$ (green, – · –), and IMEX$(6, 4; 1)$ (green, - · · -). The results shown in the top, center, and bottom panels have been obtained with meshes composed of 100, 400, and 1000 grid points, respectively. We observe that among the second-order methods, IMEX$(2, 2; 1)$ always outperforms IMEX$(2, 2; \frac{1}{2})$. Among the third-order methods, IMEX$(4, 3; 1)$ outperforms the other two. Notice also that the popular IMEX$(3, 3; 0.26)$ method is the least robust and accurate. Among the two fourth-order methods, IMEX$(5, 4; 1)$ seems to be more accurate, but IMEX$(6, 4; 1)$ seems to be more robust with respect to the CFL number. Overall, this test demonstrates that IMEX$(4, 3; 1)$ performs very well (it systematically outperforms all the other methods) and is very robust with respect to the CFL number. The second method in performance level is IMEX$(6, 4; 1)$.

5.3.2. Two-dimensional simulations. We now reproduce the test introduced by [11, 12]. The full description of the test is documented in [17, sect. 7.4]. We use $\gamma = 1.4$, $\mu = 10^{-3}$, $\lambda = 0$, $P_r := \frac{\mu c_P}{\kappa} = 0.73$ (this gives $\kappa = \frac{7}{1.46} \times 10^{-2}$). The computational domain is $D = [0, 1] \times [0, \frac{1}{2}]$. We use $\mathbb{P}_1$ Lagrange elements. The meshes are unstructured and Delaunay. They are also fitted along the segment $\{\frac{1}{2}\} \times [0, \frac{1}{2}]$ to approximate the initial data as best as possible. The time stepping is done with the method IMEX$(4, 3; 1)$ at CFL $= 1.5$. The time step is computed by means of the definition (5.7) with $s = 4$.

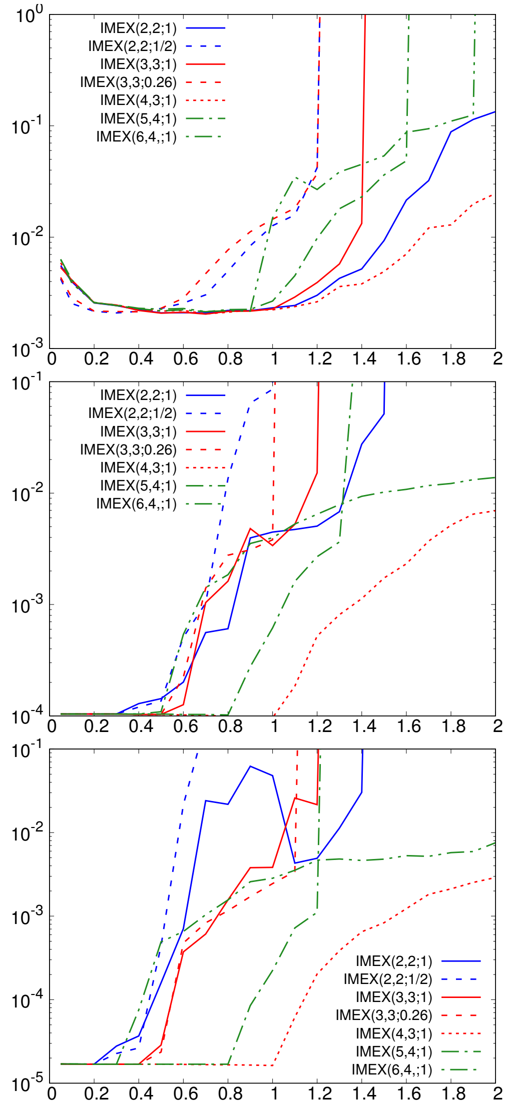

FIG. 5.2. *Becker's solution at* $T=3$ *with* $\mu=10^{-2}$. L^1*-error (see* (5.10)*) as a function of the CFL number. Top panel:* 100 *grid points. Middle panel:* 400 *grid points. Bottom panel:* 1000 *grid points. (Color available online.)*

Denoting $g(\boldsymbol{x})=\|\nabla\rho_h(\boldsymbol{x})\|_{\ell^2}$, $g_{\min}=\min_{\boldsymbol{x}\in D} g(\boldsymbol{x})$, $g_{\max}=\max_{\boldsymbol{x}\in D} g(\boldsymbol{x})$, we show in Figure 5.3 the quantity $e^{-10\frac{g-g_{\min}}{g_{\max}-g_{\min}}}$. This representation (similar to Schlieren photography) amplifies the contrast in the density field. We show in the top, center, and bottom panels of the figure the density field at $t=1$ for meshes composed of 328,253, 572,301 and 761,879 grid points, respectively. We observe that the results are consistent with those reported in [17, sect. 7.4].

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 demonstrate that the

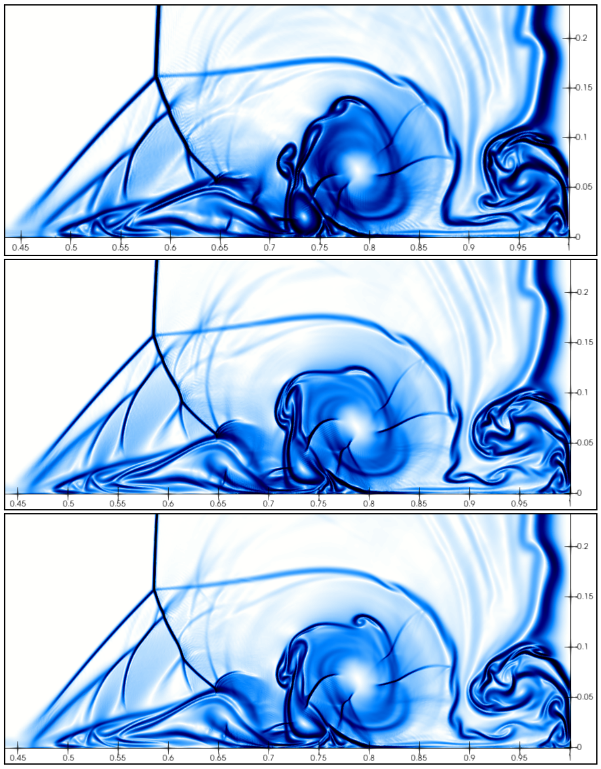

FIG. 5.3. *Two-dimensional shocktube test. Density at* $t=1$ *with* $\mu=10^{-3}$. *Time stepping done with* IMEX(4, 3; 1) *and CFL=* 1.5. *From top to bottom, meshes with increasing refinement level:* 328,253 *grid points;* 572,301 *grid points;* 761,879 *grid points.*

new IMEX methods proposed in the paper behave as predicted by the theory. All the methods tested are invariant-domain preserving and deliver the expected accuracy. A natural extension of this work is to show how the present methodology can be used to solve the compressible Navier–Stokes equations with temperature-dependent properties and nonideal thermodynamics.

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., 3 (1995), pp. 797–823, https://doi.org/10.1137/0732037.

[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 (1997), pp. 151–167.

[3] R. Becker, *Stoßwelle und Detonation*, Z. Phys., 8 (1922), pp. 321–362.

[4] 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 (1997), pp. 172–186.

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

[6] 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 (2012), pp. 681–707.

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

[8] 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 (1991), pp. 169–210.

[9] M. Crouzeix, *Sur l'approximation des équations différentielles linéaires par des méthodes de Runge-Kutta*, Ph.D. thesis, Univ. Paris 6, France, 1975.

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

[11] V. Daru and C. Tenaud, *Evaluation of TVD high resolution schemes for unsteady viscous shocked flows*, Comput. & Fluids, 30 (2001), pp. 89–113.

[12] V. Daru and C. Tenaud, *Numerical simulation of the viscous shock tube problem by using a high resolution monotonicity-preserving scheme*, Comput. & Fluids, 38 (2009), pp. 664–676.

[13] A. Ern and J.-L. Guermond, *Invariant-domain-preserving high-order time stepping: I. Explicit Runge–Kutta schemes*, SIAM J. Sci. Comput., 44 (2022), pp. A3366–A3392, https://doi.org/10.1137/21M145793X.

[14] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489, https://doi.org/10.1137/16M1074291.

[15] 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 (2018), pp. A3211–A3239, https://doi.org/10.1137/17M1149961.

[16] 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 (2019), pp. 143–175.

[17] 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 (2021), 113608.

[18] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-algebraic Problems*, 2nd revised ed., Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, 2010.

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

[20] B. M. Johnson, *Analytical shock solutions at large and small Prandtl number*, J. Fluid Mech., 726 (2013), R4.

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

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

[23] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558.

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

[25] D. Kuzmin, M. Quezada de Luna, D. I. Ketcheson, and J. Grüll, *Bound-preserving flux limiting for high-order explicit Runge-Kutta time discretizations of hyperbolic conservation laws*, J. Sci. Comput., 91 (2022), 21.

[26] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes.* I, SIAM J. Numer. Anal., 33 (1996), pp. 760–779, https://doi.org/10.1137/0733038.

[27] C. Lohmann and D. Kuzmin, *Synchronized flux limiting for gas dynamics variables*, J. Comput. Phys., 326 (2016), pp. 973–990.

[28] S. P. Nørsett, *Semi Explicit Runge-Kutta Methods*, Technical report 6/74, Univ. Trondheim, 1974.

[29] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM J. Numer. Anal., 21 (1984), pp. 955–984, https://doi.org/10.1137/0721060.

[30] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*, in Recent Trends in Numerical Analysis, Adv. Theory Comput. Math. 3, Nova Science, Huntington, NY, 2001, pp. 269–288.

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

[32] M. Quezada de Luna and D. I. Ketcheson, *Maximum principle preserving space and time flux limiting for diagonally implicit Runge-Kutta discretizations of scalar convection-diffusion equations*, J. Sci. Comput., 92 (2022), 102.

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

[34] C. Schär and P. K. Smolarkiewicz, *A synchronous and iterative flux-correction formalism for coupled transport equations*, J. Comput. Phys., 128 (1996), pp. 101–120.

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

[36] J. M. Varah, *Stability restrictions on second order, three level finite difference schemes for parabolic equations*, SIAM J. Numer. Anal., 17 (1980), pp. 300–309, https://doi.org/10.1137/0717025.

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

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

[39] 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 (2011), pp. 2752–2776.

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